EDA CASE STUDY

Group Name:

1 Shristi Anand

2 Swathi Kommana

3 Devasena

4 Umesh Masuram

Introduction

Objective:

- Identification of Loan Applicant traits that tend to ‘default’ paying back
- Understand the ‘Driving Factors’or ‘Driver Variables’ behind Loan Default phenomena
- Gramener may choose to utilize this knowledge for its portfolio and risk assessment of new loan applicants

Assumptions :

•The variables like recoveries, total_pymnt, total_pymnt_inv, total_rec_prncp etc.. which normally get captured only after a loan is accepted, will not be available at the time of a new loan application. So these type of variables can be removed from the dataset.

•Since bankruptcy filings, tax liens and judgments are the three kinds of public records that appears on a credit report, this information should already be captured in column pub_rec which contains derogatory public records.

•Purpose and title have redundant information.

•Emp_title column has so much discrepancies in its values(e.g. The same employer name is mentioned in various formats) .Also as it has many unique values it would not give any useful insights about the pattern for loan defaulting.

Data Cleaning and Manipulation

Dealing with Missing values :

•In this dataset 54 variables contain all the observations as NA so, they are removed.

•14 columns have more than 70% of 0 so they are removed.

•Removing Columns with 90% or more zeros and NA’s

•4 columns from remaining has same data in column hence they were removed.

•Date is converted into standard format and % is removed from columns wherever required.

•Removed the variables having too many levels like title.

•Remove 'xx' from ‘zip_code’ and Create Address column by combining values in ‘zip_code’ and 'addr_state’.

•Imputed employee length with mean and rounded to 2 decimal places

•funded_amnt_inv, installment, int_rate Round to 2 decimal places

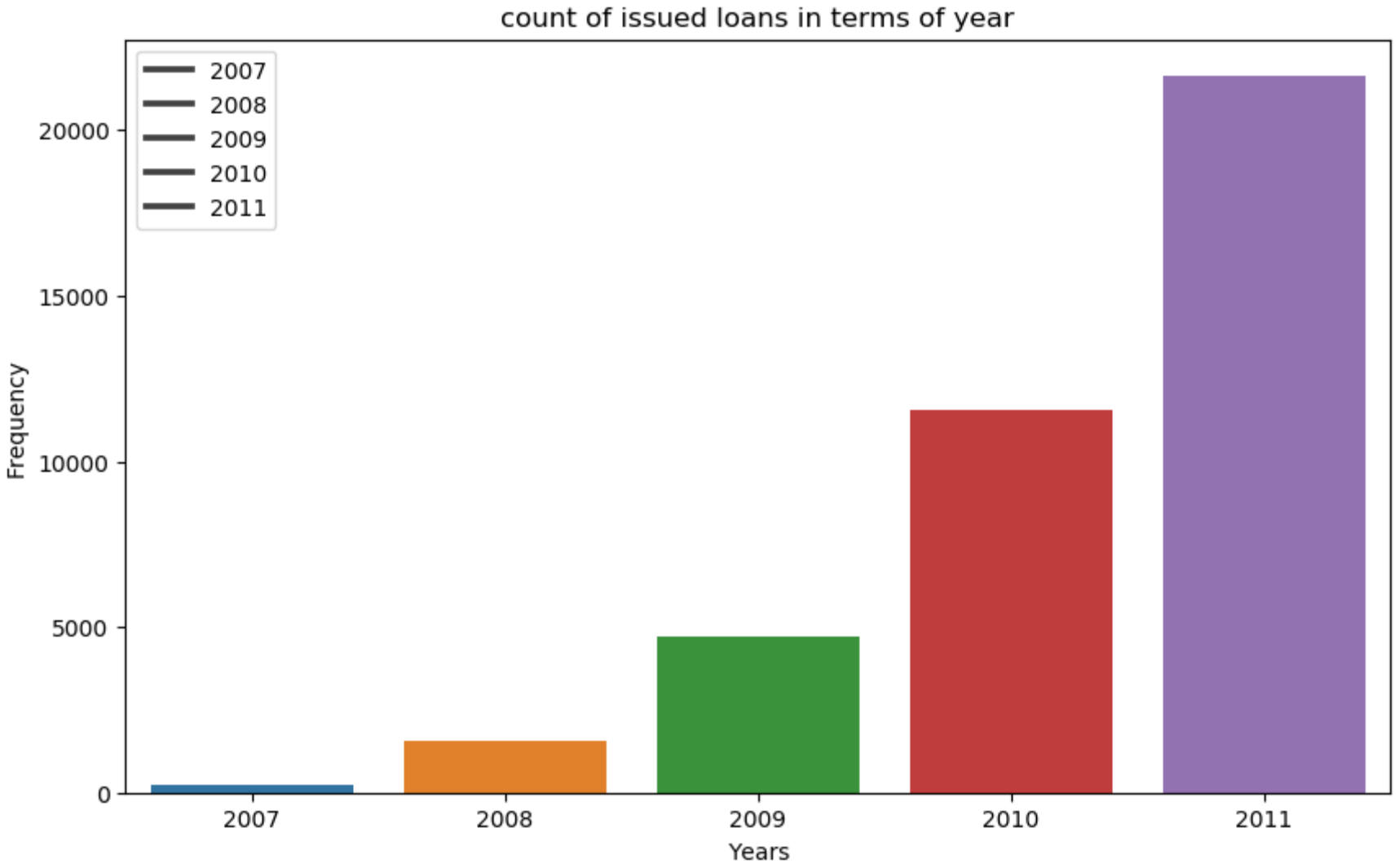
.

Top 6 Deciding Parameters for Loan Defaulter

1. Loan Amount
2. Grade
3. Purpose of Loan
4. State
5. Home Ownership
6. Verification Status

Subsequent slides are the visualization of above variables.

Analysis of loans Issued per year



Inference: The Number of issued loans increases year by year.

Percentage increase for year 2007 = 0.0 %

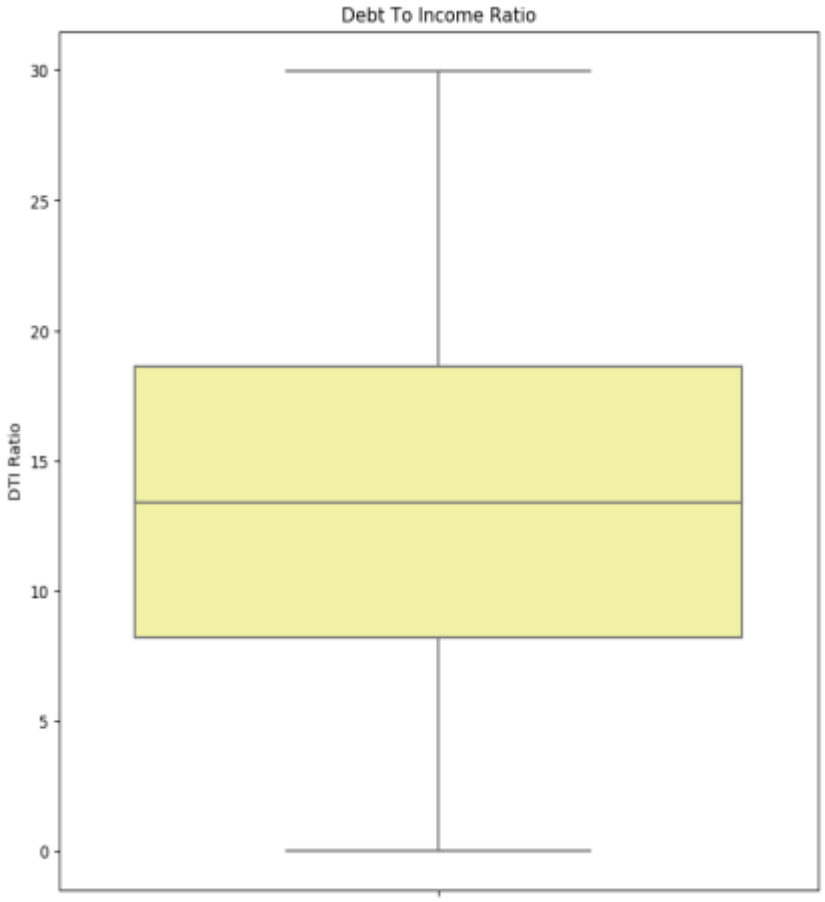
Percentage increase for year 2008 = 527.0 %

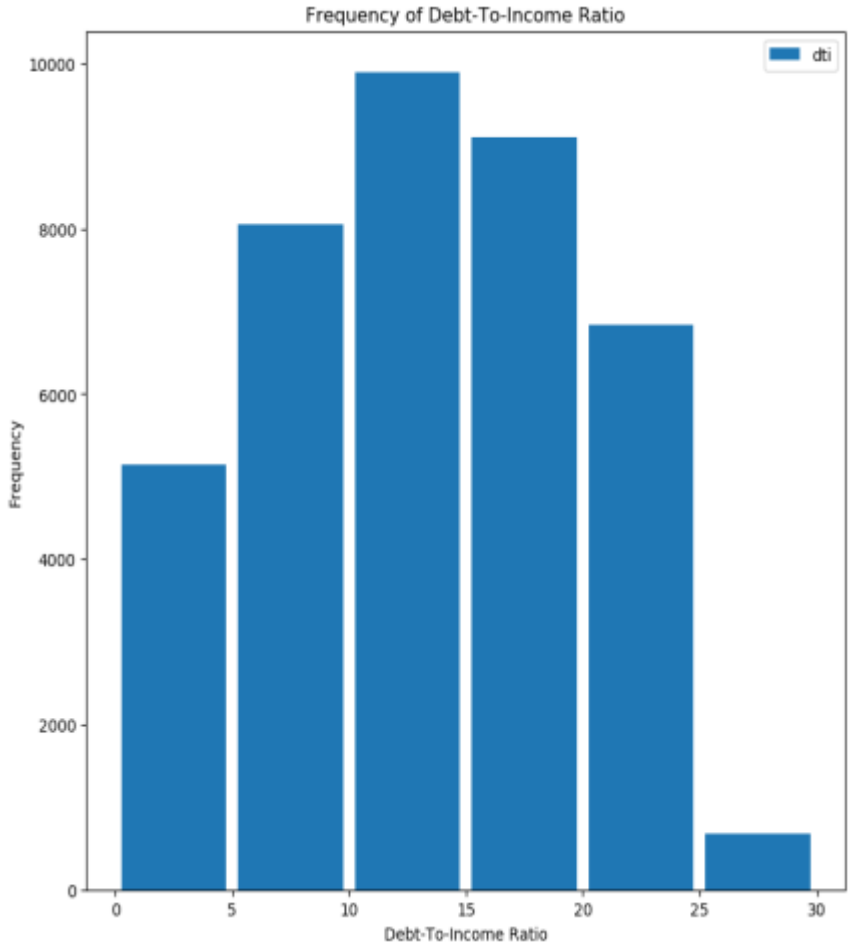
Percentage increase for year 2009 = 202.0 %

Percentage increase for year 2010 = 145.0 %

Percentage increase for year 2011 = 88.0 %

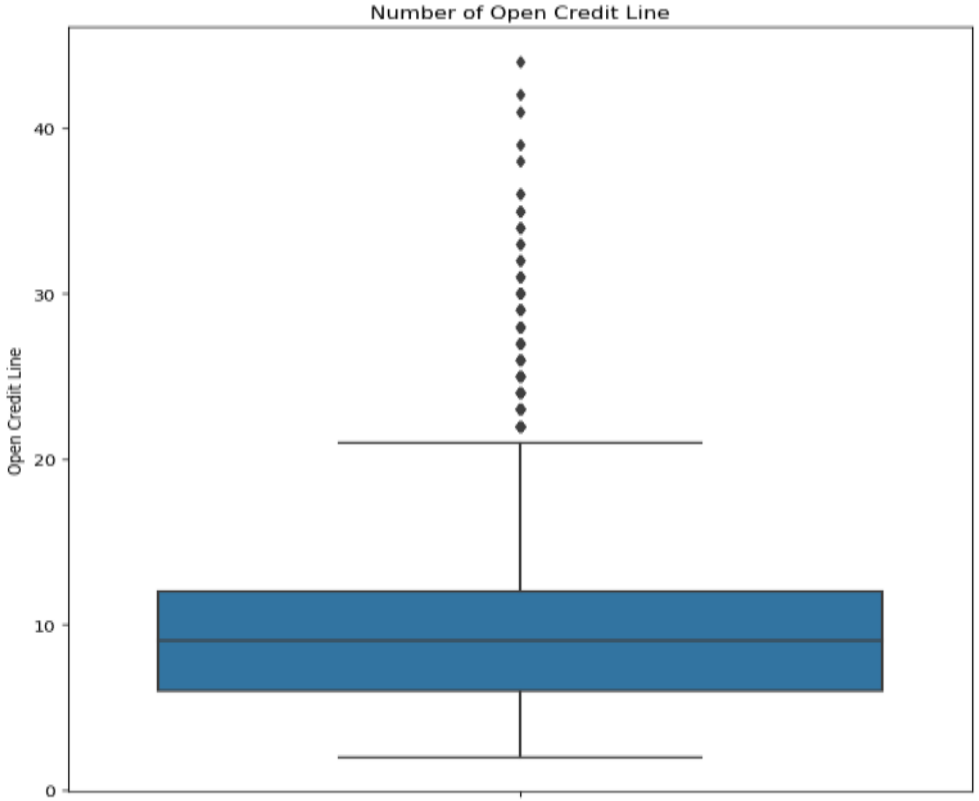
Analysis of Debt-To-Income Ratio(dti)

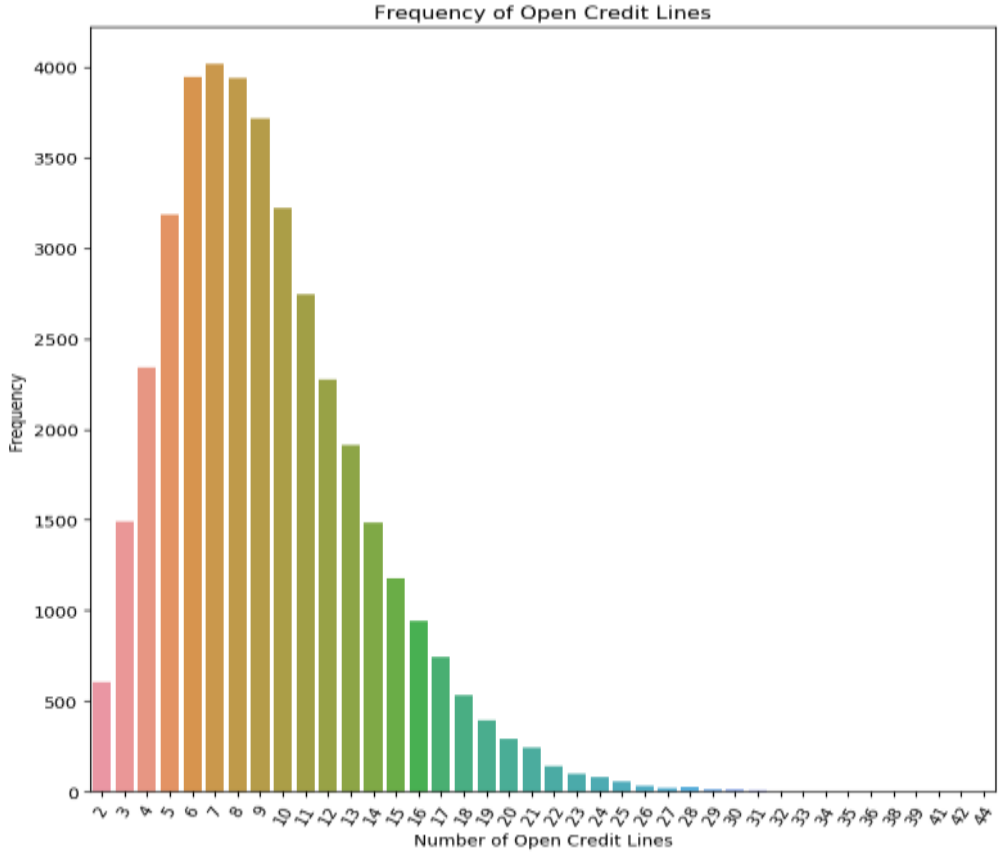




Inferences : dti values are equally distributed along the median between 25-75 percentile with max value, Approximately at 30, minimum at 0 and maximum borrowers given loan having DTI value between 10-15%

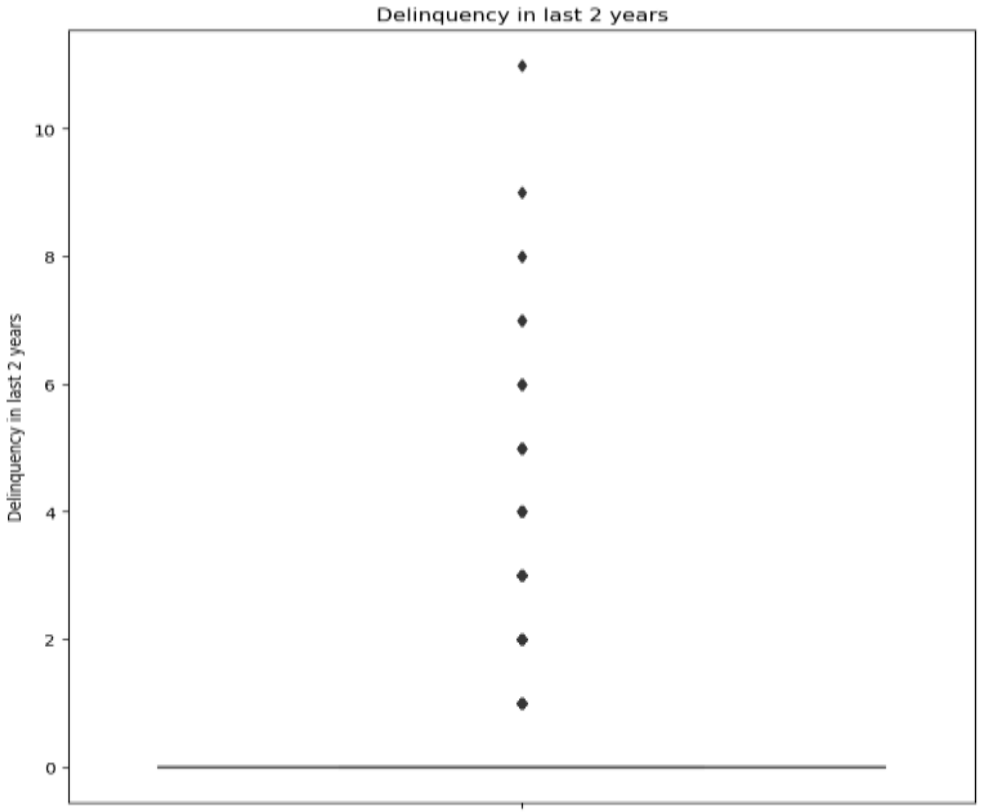
Analysis of Number of open credit lines in the borrower's credit file(open_acc)

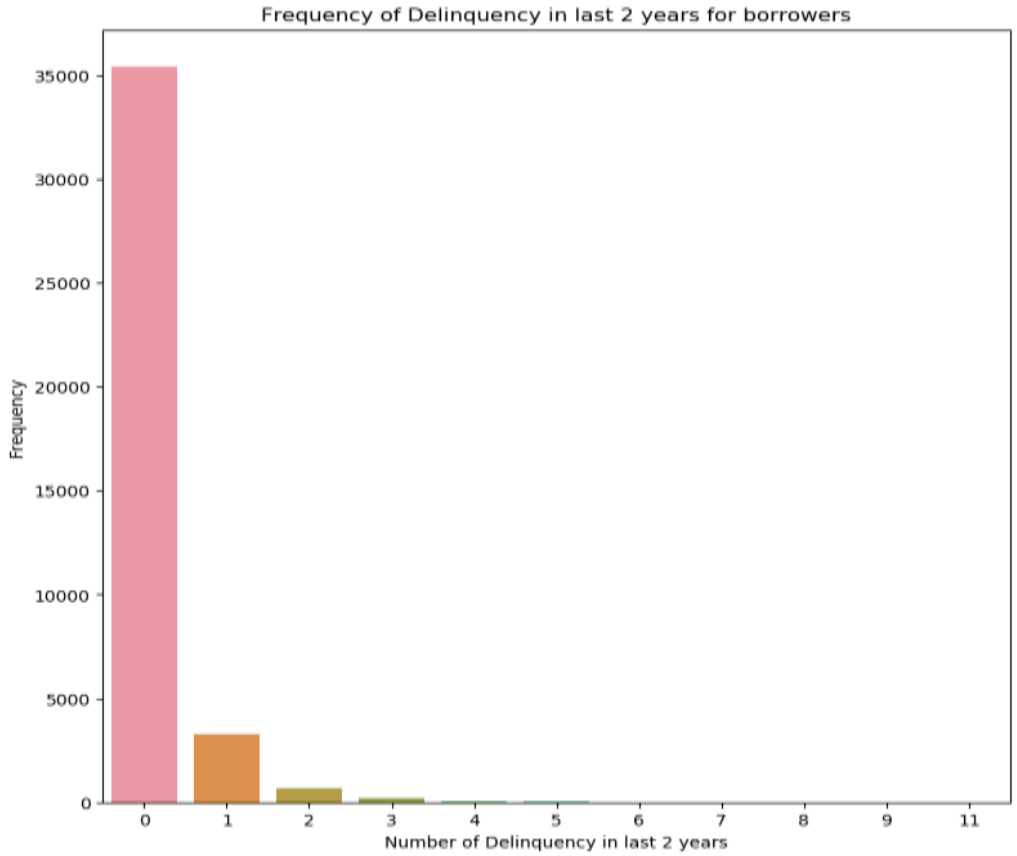




Inference:
open_acc values seems equally distributed along the median(9) for 25-75 percentile with some outliers.

Analysis of Delinquency in past 2 Years(delinq_2yrs)

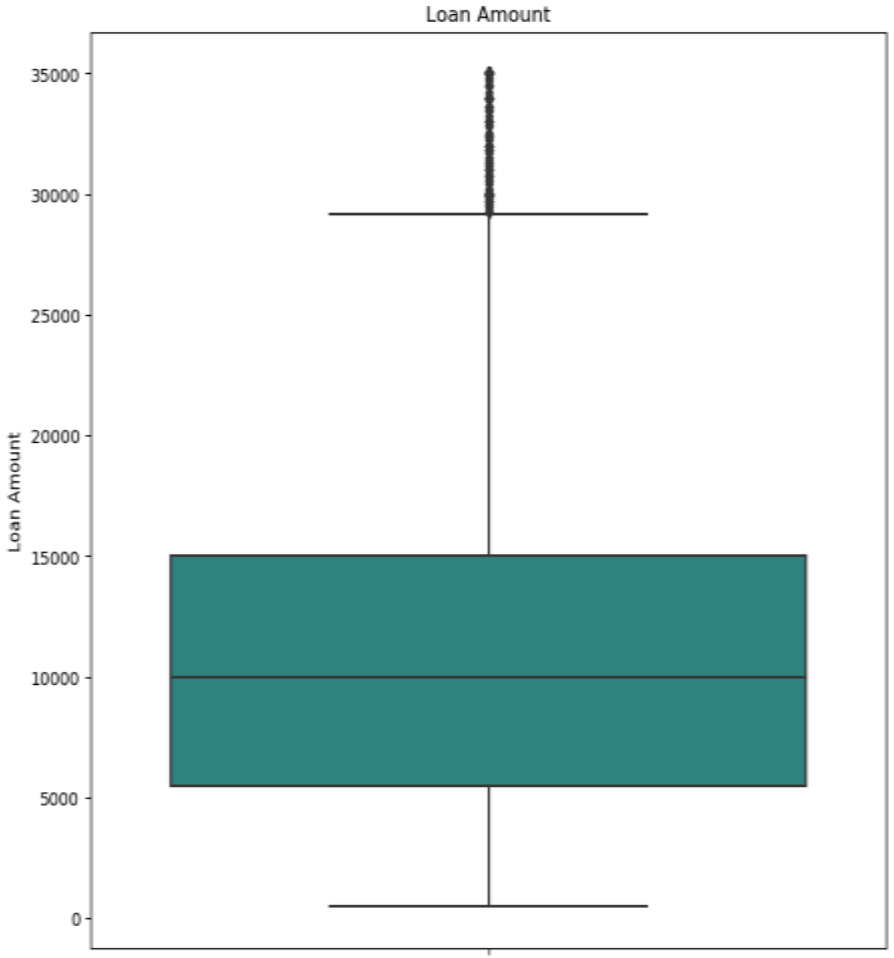


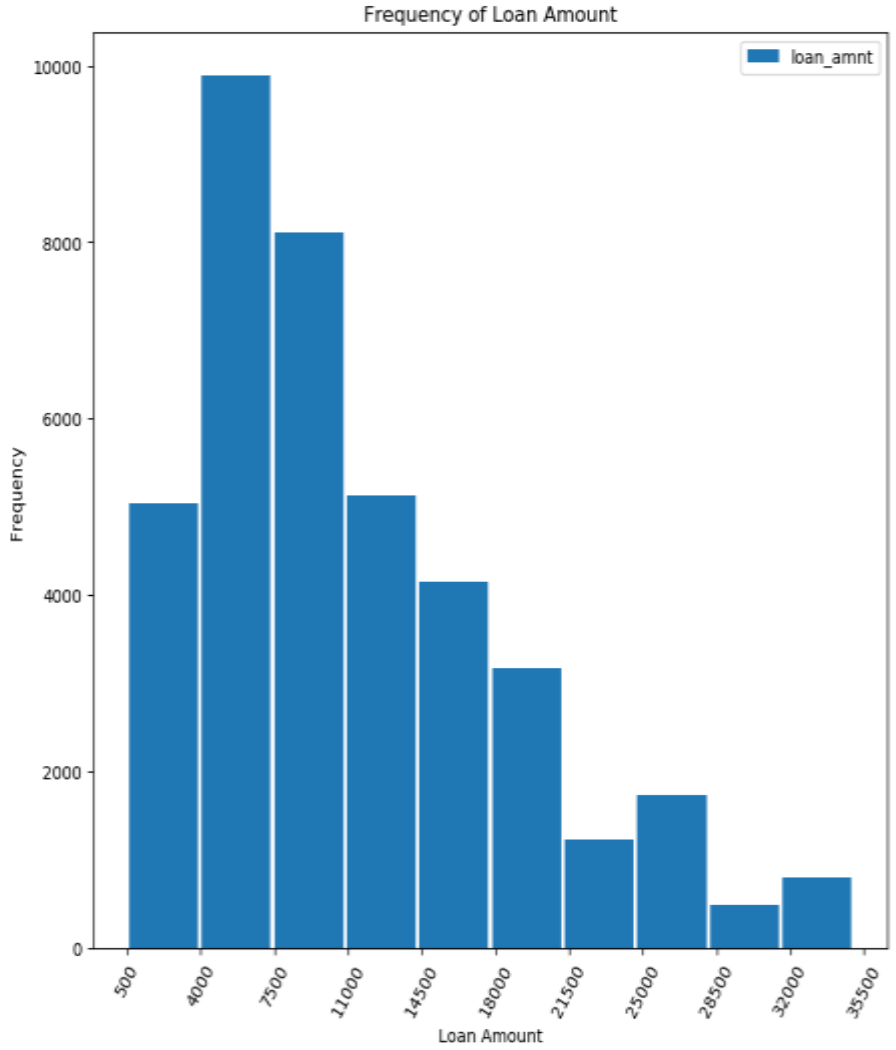


Inference:

Most of the delinq_2yrs values are 0 and hence highest frequency is for value 0.
There are some outliers and the max value is 11.

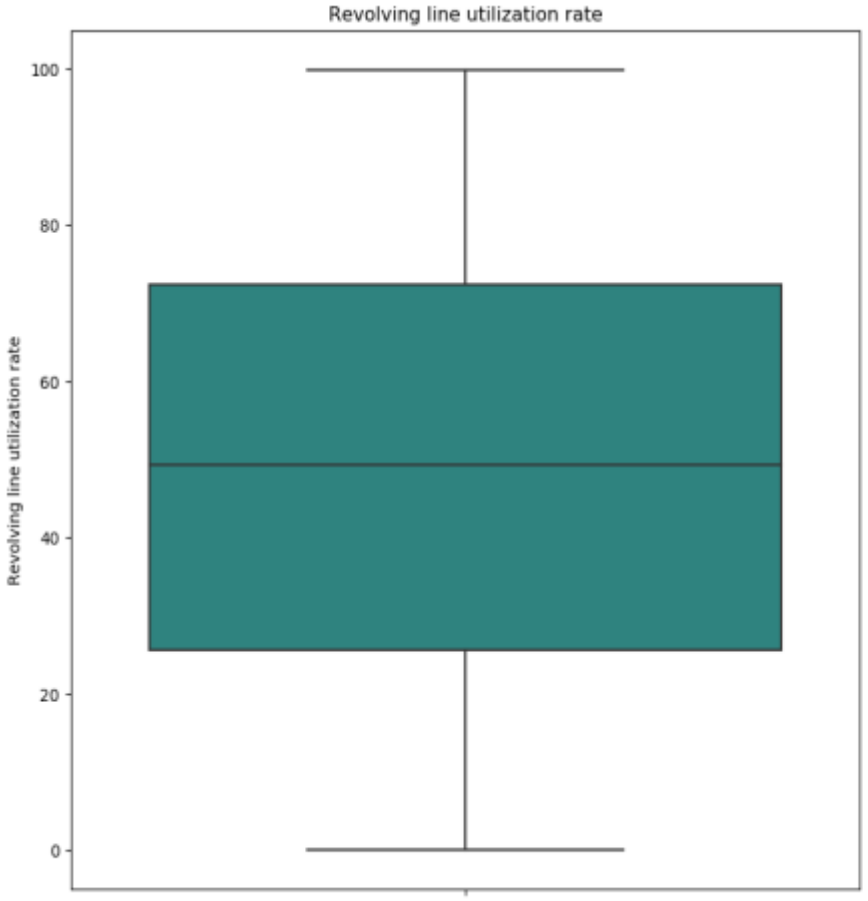
Analysis of quantitative variable loan_amnt

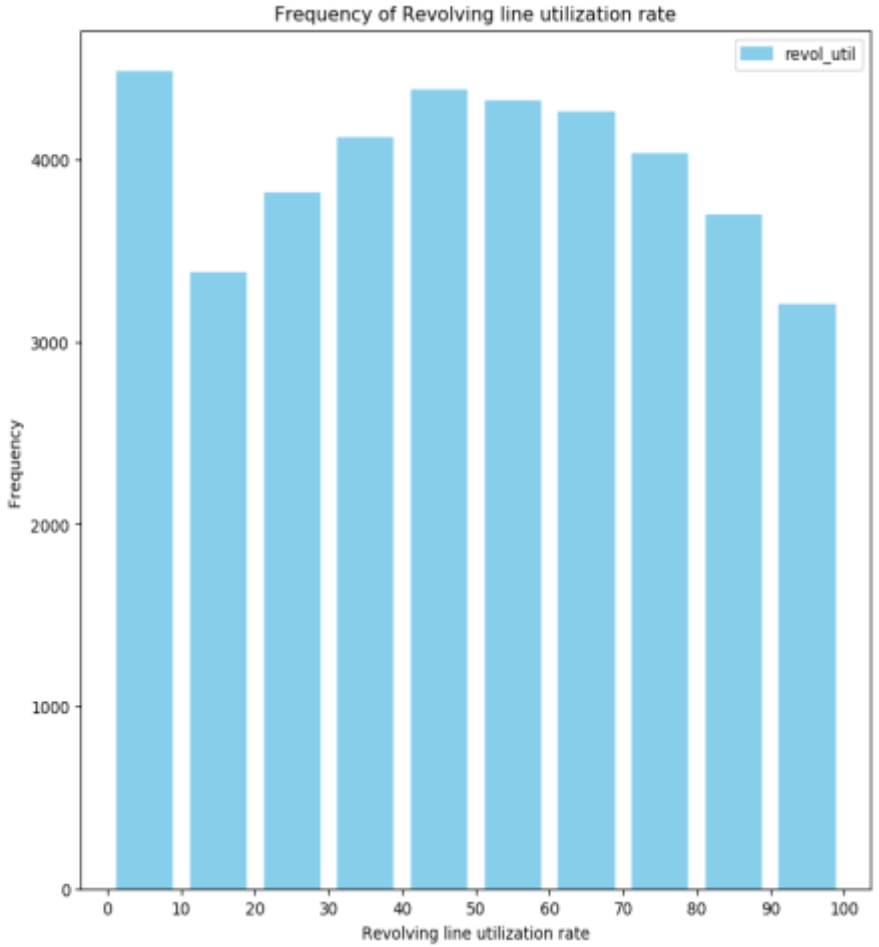




Inference:
The 'Loan Amount' most frequently taken is between $4000 to $7500. But there are clearly some outliers above 98 percentile.

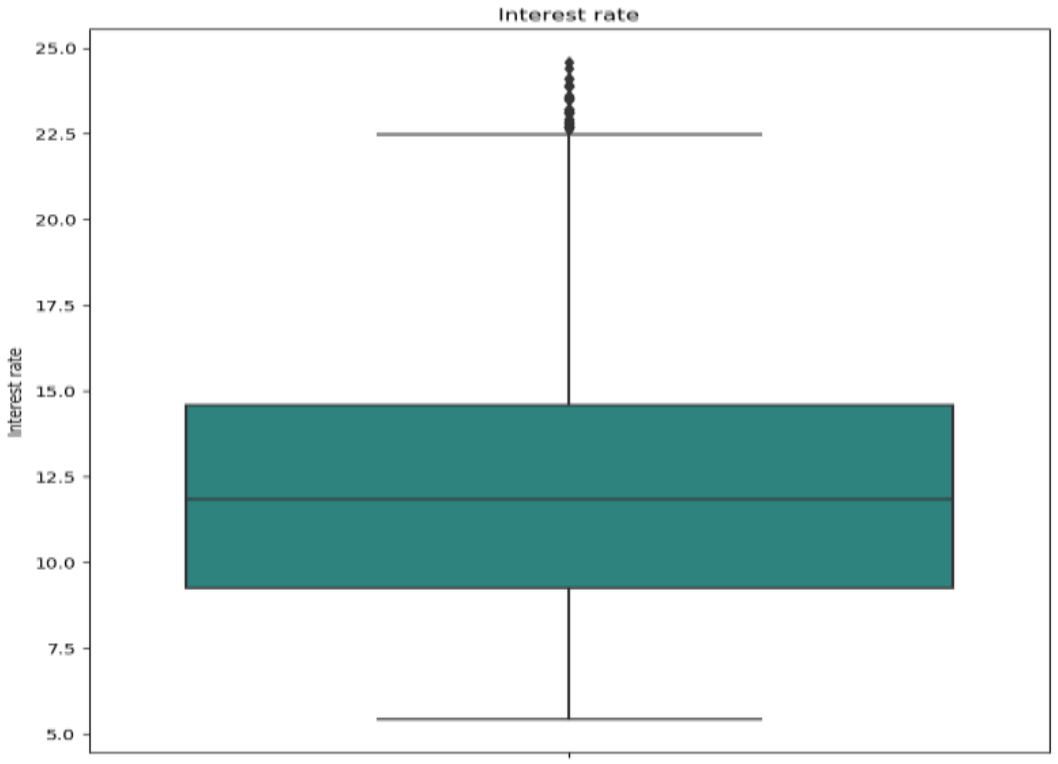
Analysis of Revolving line utilization rate(revol_util)

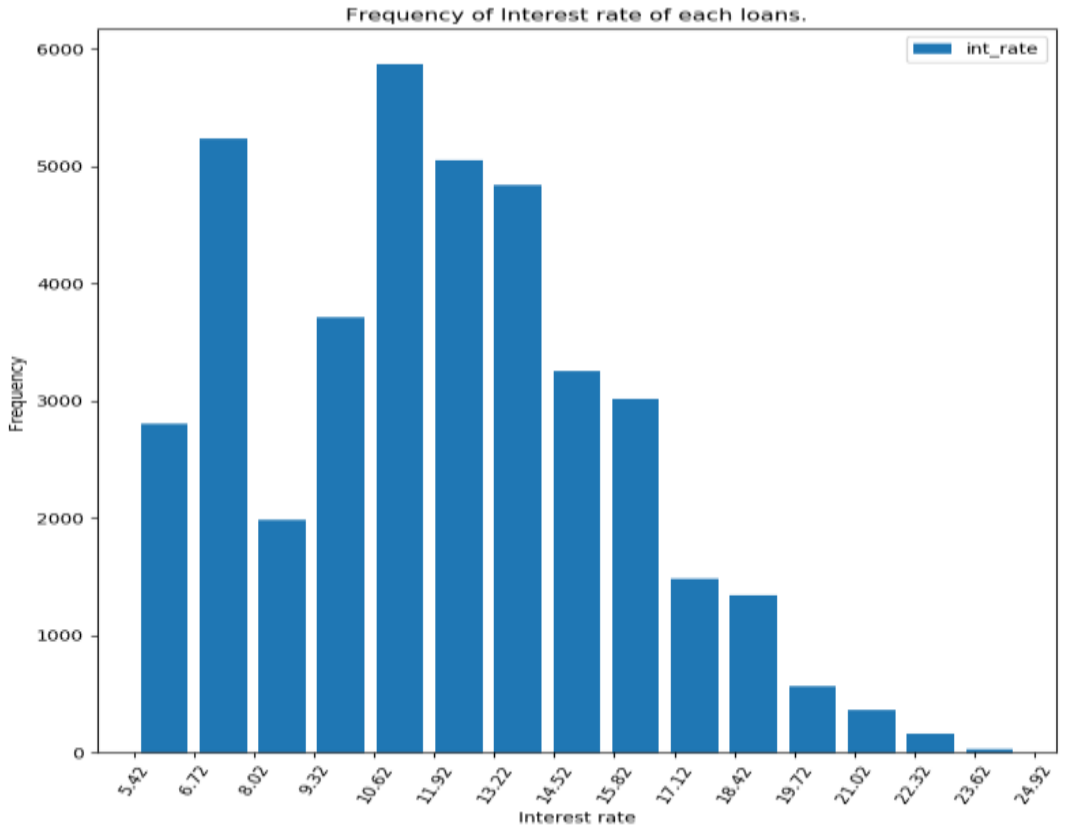




Inference:
'revol_util' values are equally distributed along the median between 25-75 percentile with max value near 100, and maximum borrowers having revolving credit utilisation rate in between 0-10%.

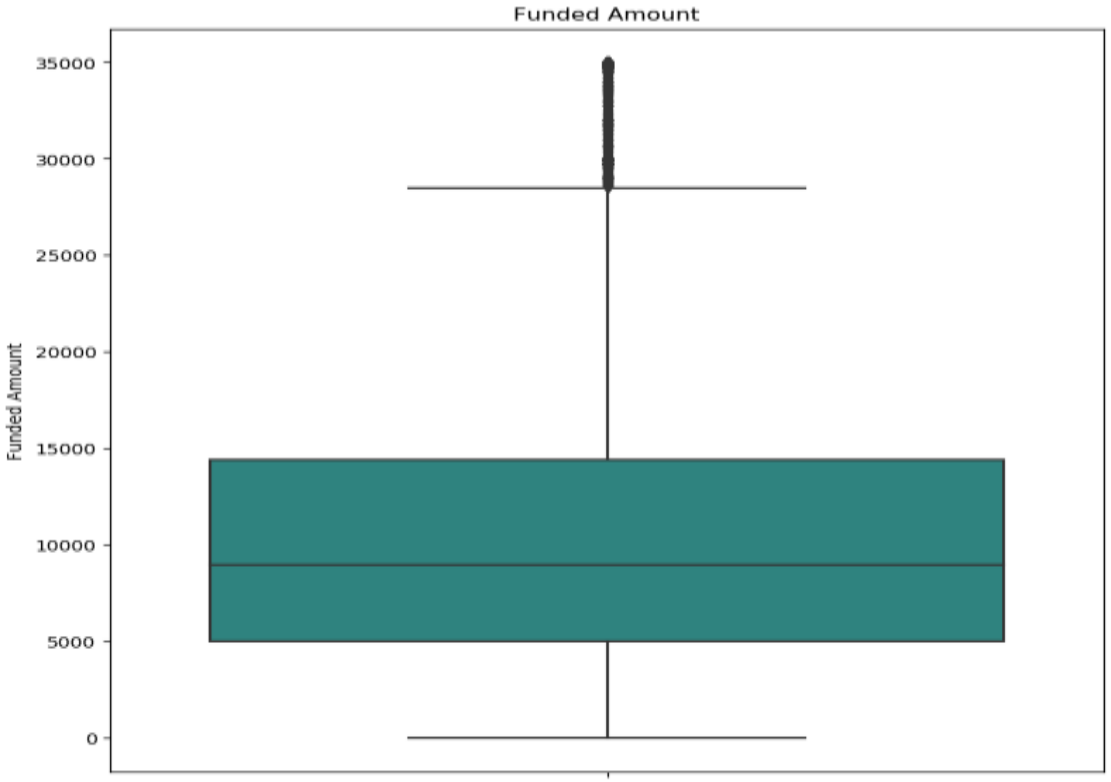
Analysis of quantitative variable int_rate

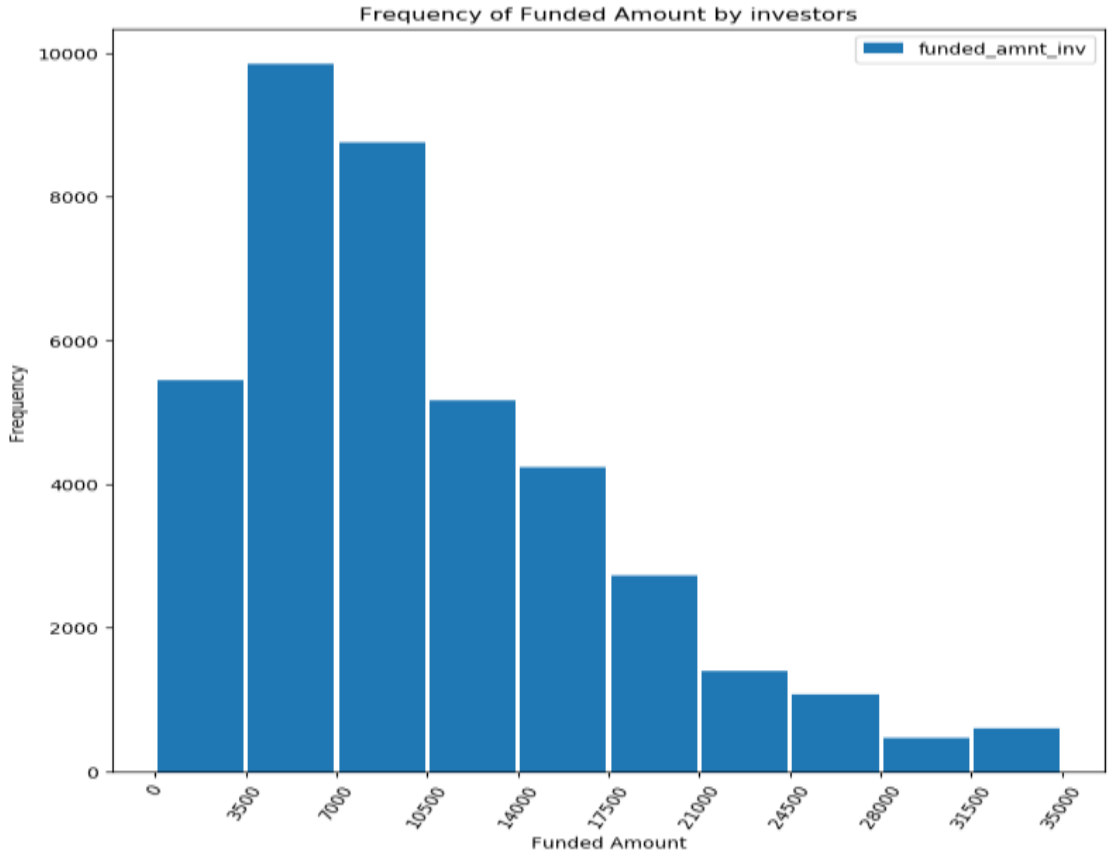




Inference: From the 'Interest rate' most frequent Interest rate is between 10.6% to 11.9% . But there are clearly some outliers above 98 percentile.

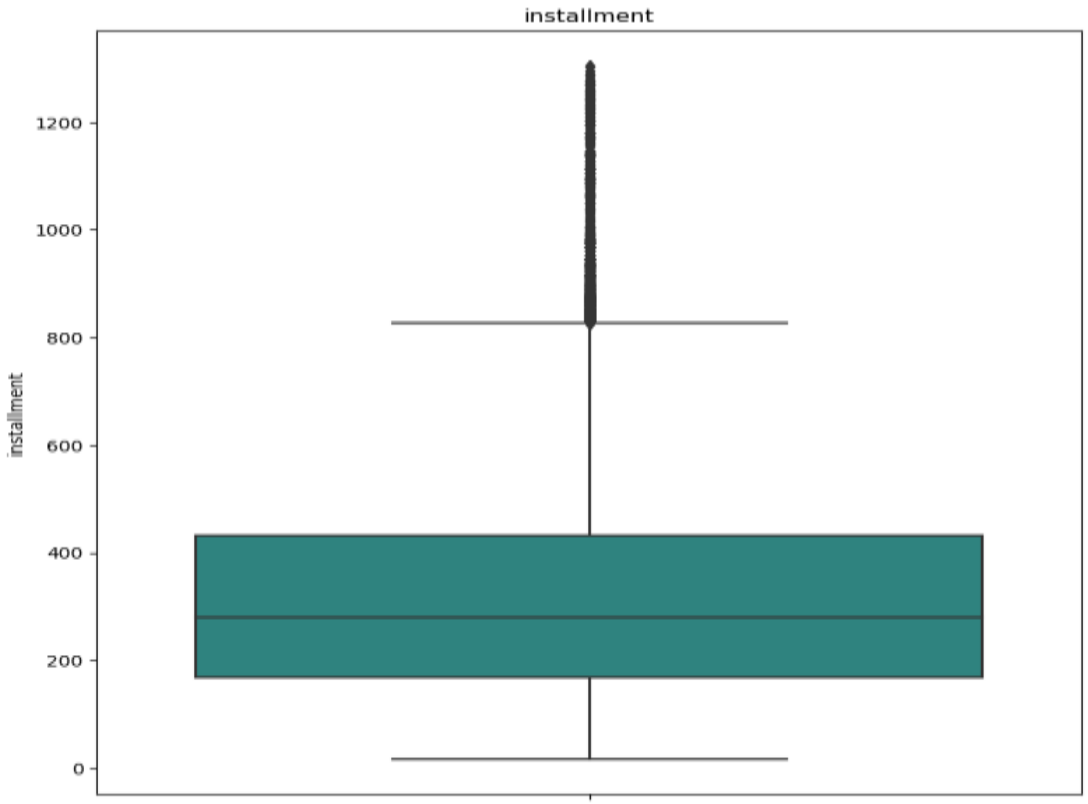
Analysis of quantitative variable funded_amnt_inv

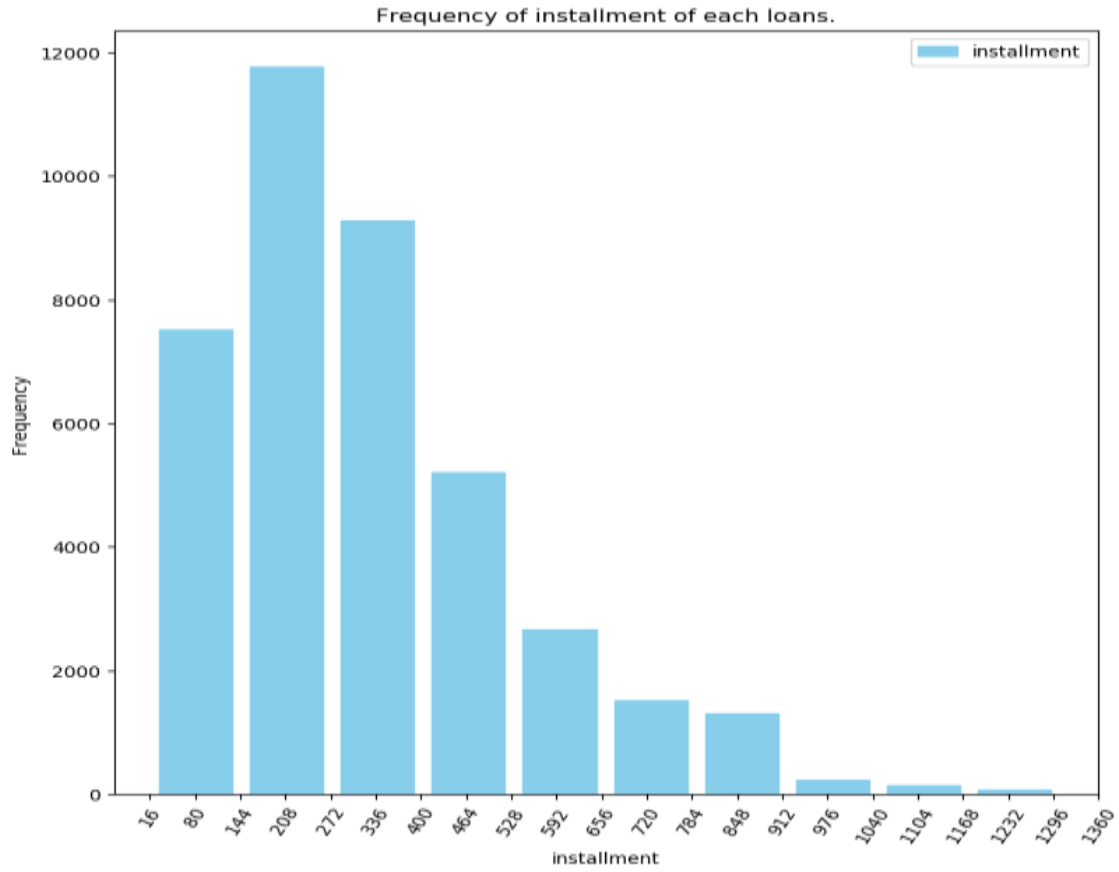




Inference: From the 'Funded Amount investors' most frequently invested is between $3500 to $7000 .
But there are clearly some outliers above 98 percentile.

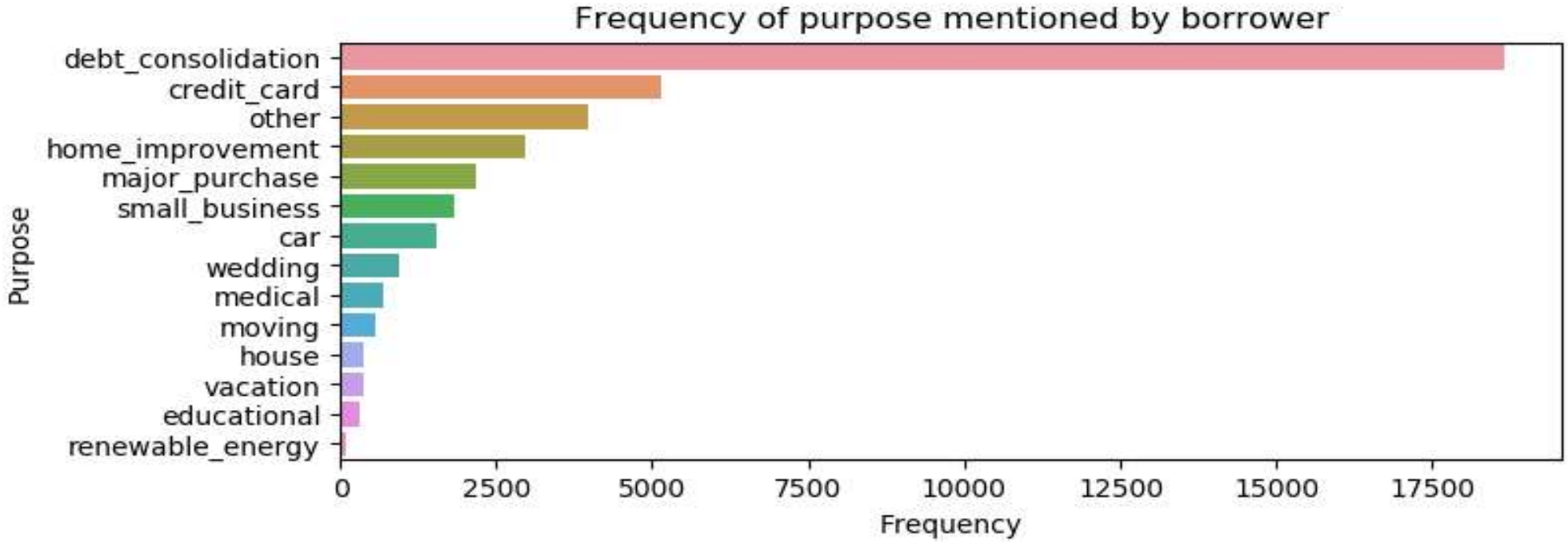
Analysis of installment





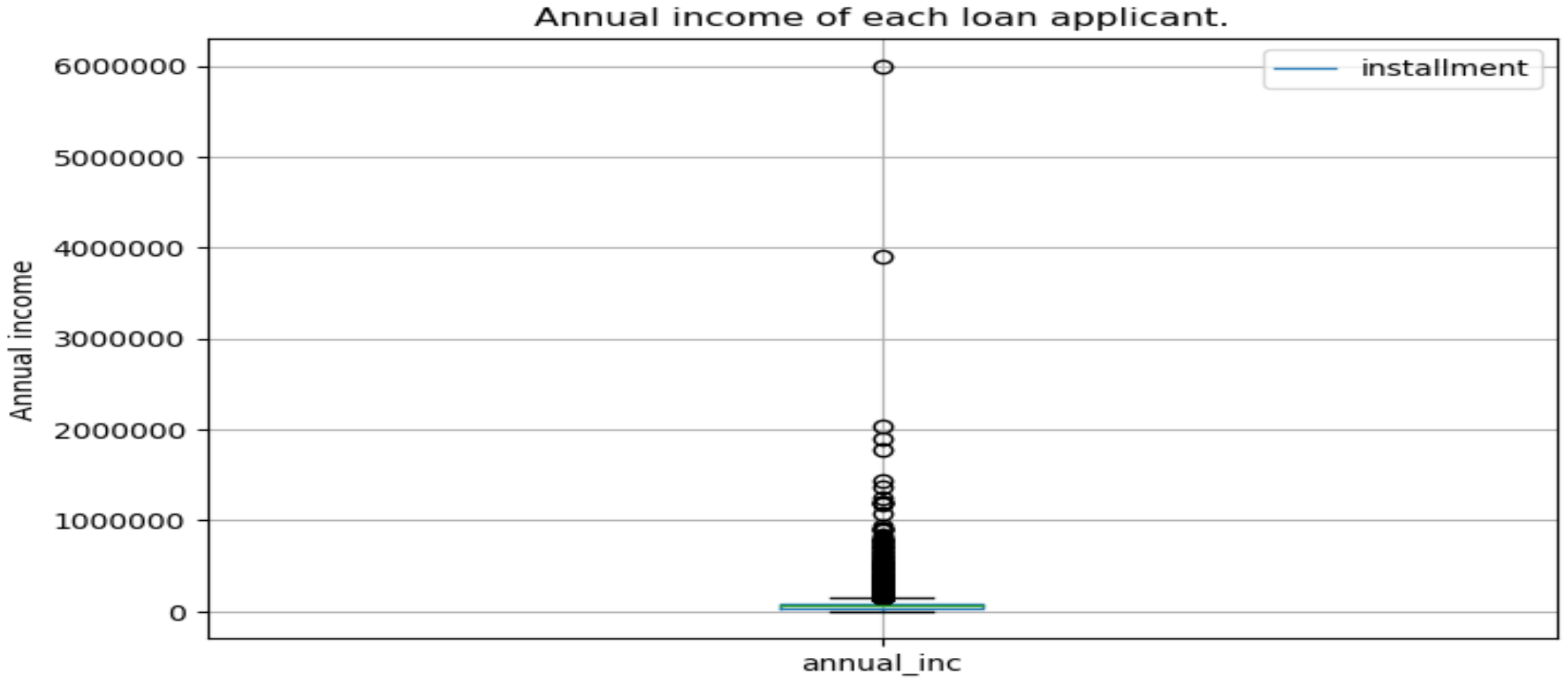
Inference: From the 'installment' column most frequent installment is between 144 to 270 . But there are clearly some outliers above 98 percentile.

Analysis of purpose for taking loan given by the borrower



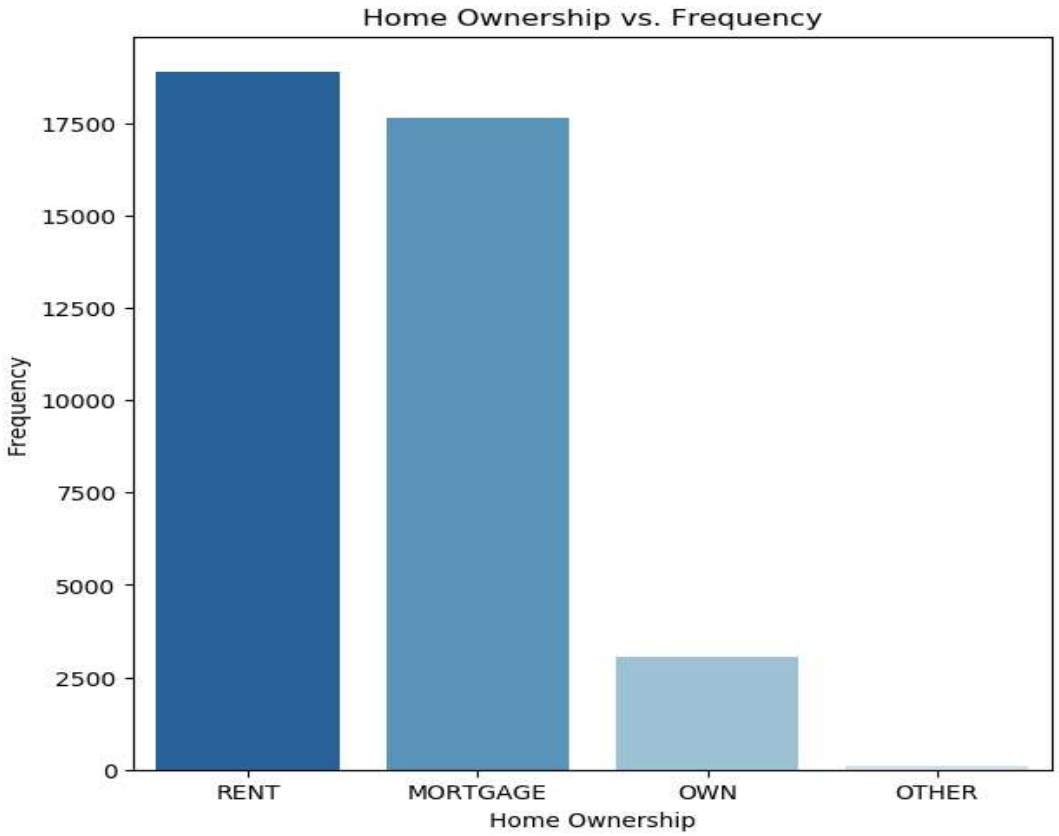
Inference: Most of the loans are taken for debt Consolidation.

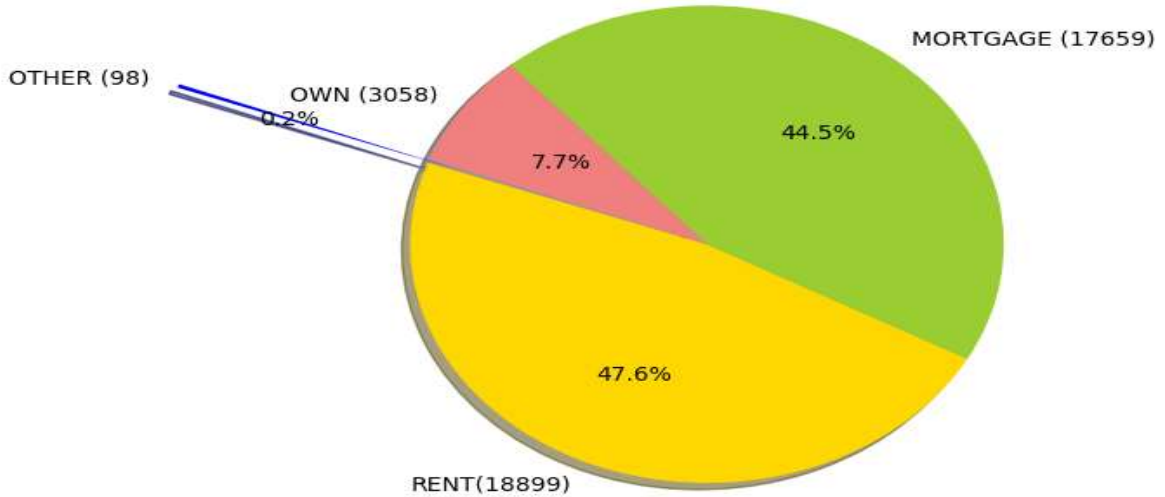
Analysis of annual income variable



Inference: From the 'Annual income', there are outliers due to which the visualization is not very effective

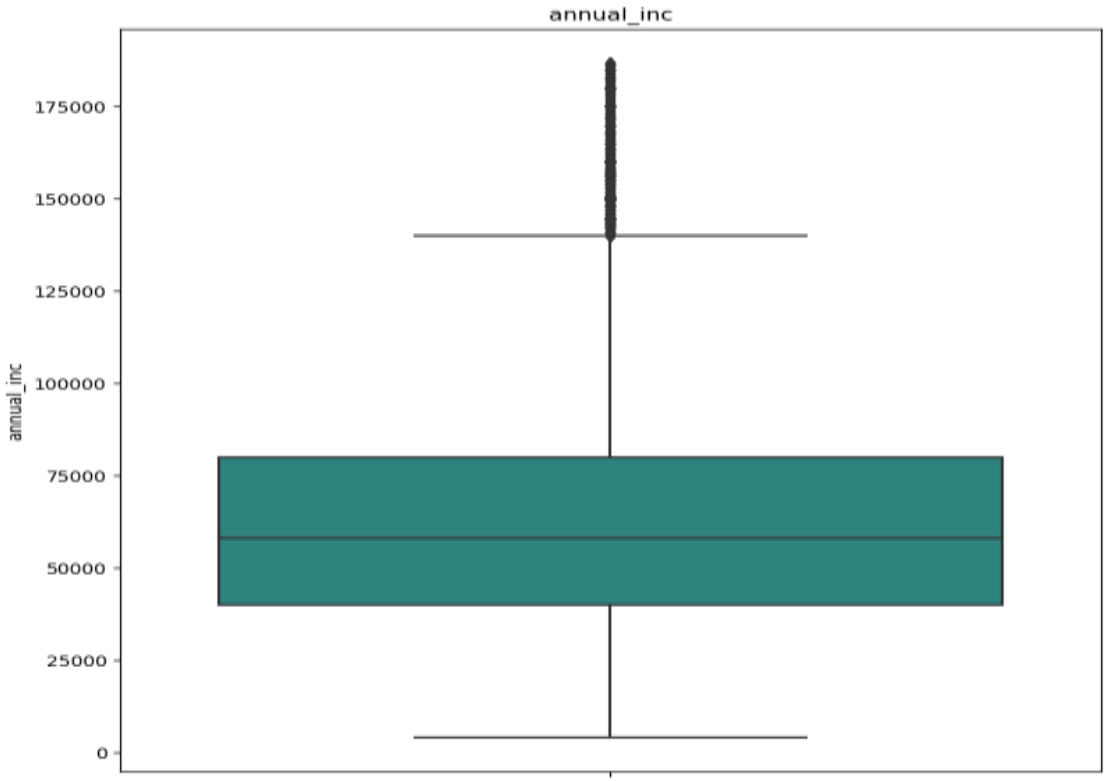
Analysis of Home Owner Ship variable

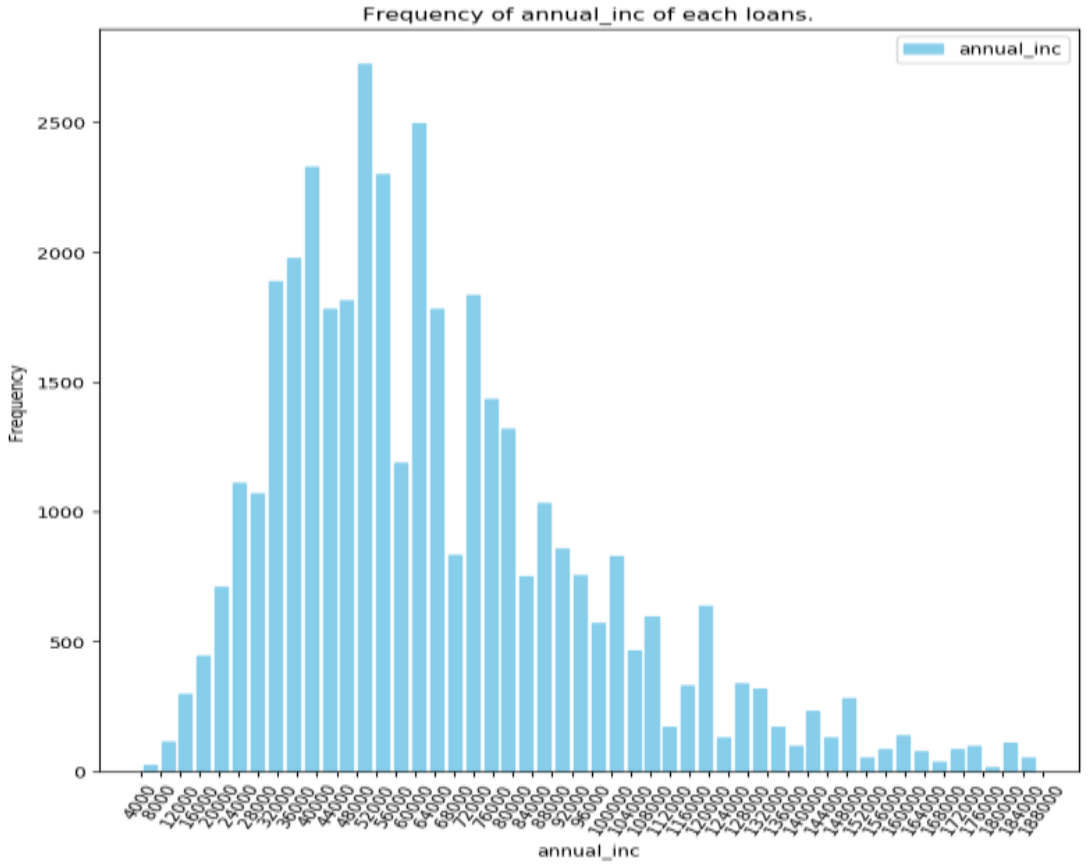




Inference: People who are staying on rented house are taking most of the loans, Followed by people staying in a mortgaged home.

Plotting annual income after removing outliers

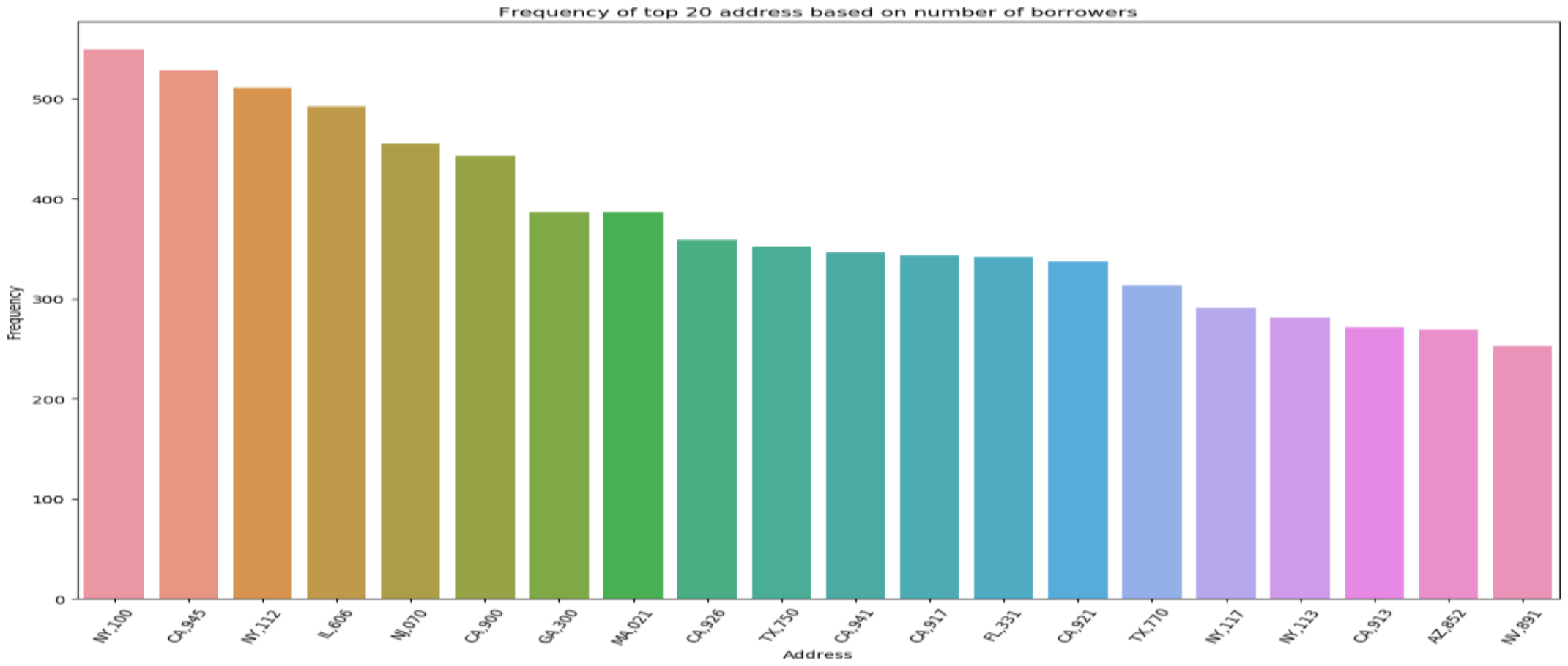




Inference:

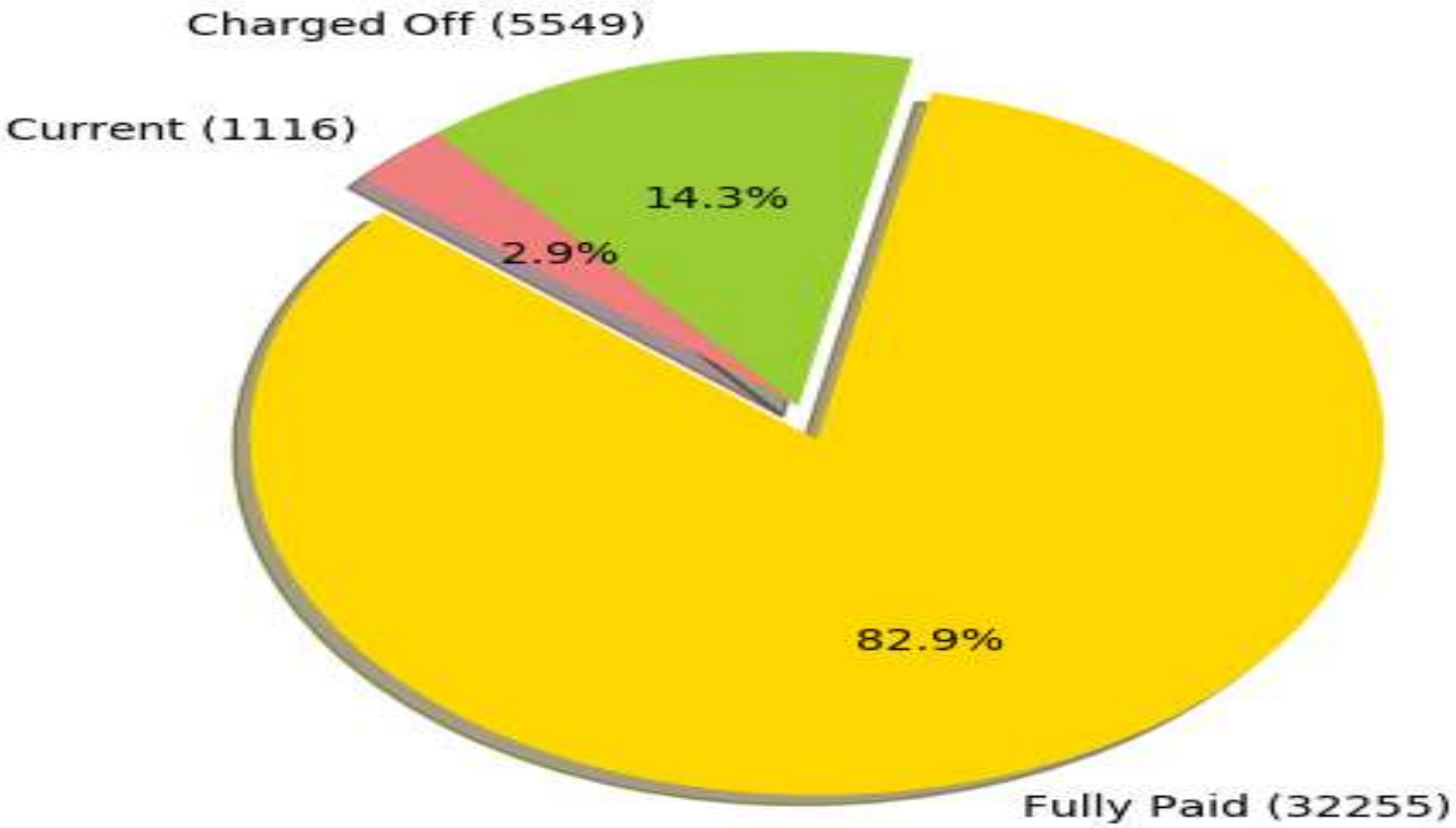
After removing the outliers the 'annual_inc' column, most frequent income range is between $46000 to $52000.

Univaritate analysis of address(combination of zipcode and state)



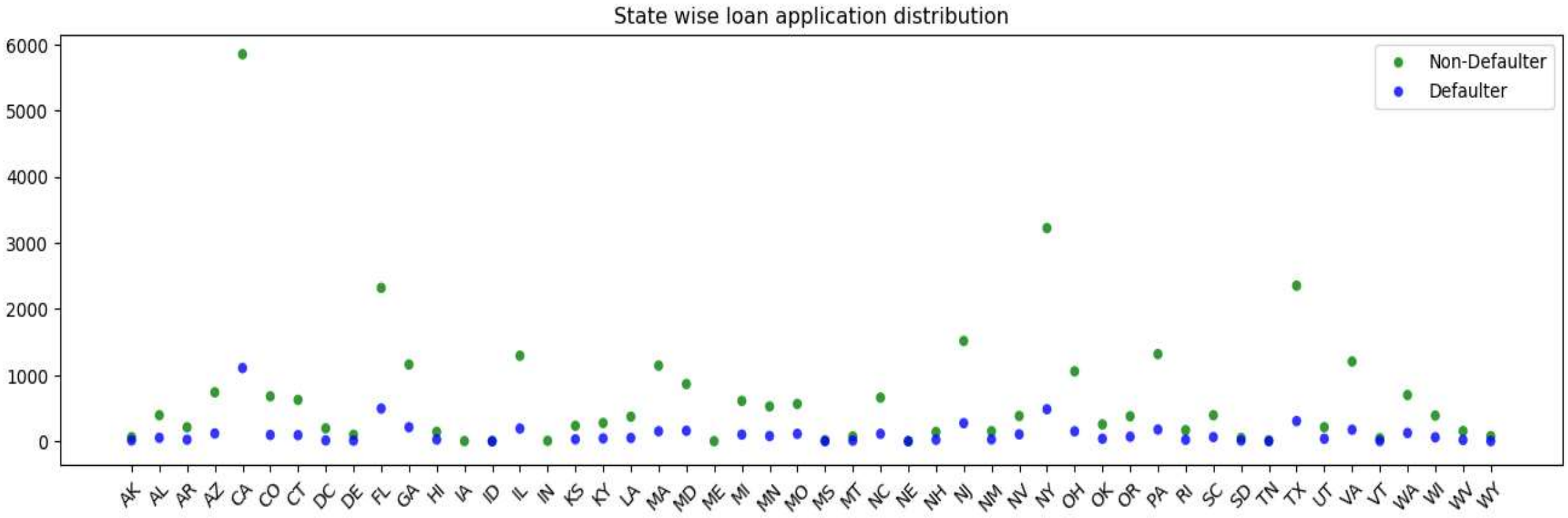
Inference: Maximum number of borrowers are from address NY,100 of loans to whom loans are given

Analysis of loan status



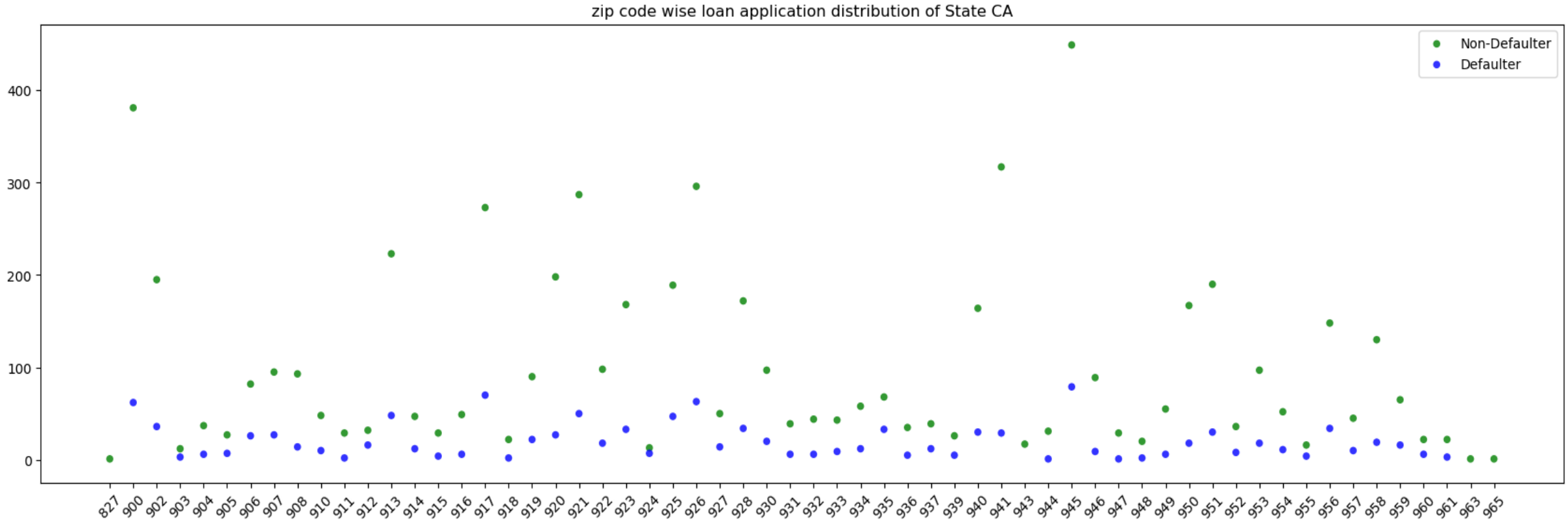
Inference: Most of the loans in the data set are fully paid and only a small percentage of loans are current loans.

Segmented Analysis of state wise loan_status



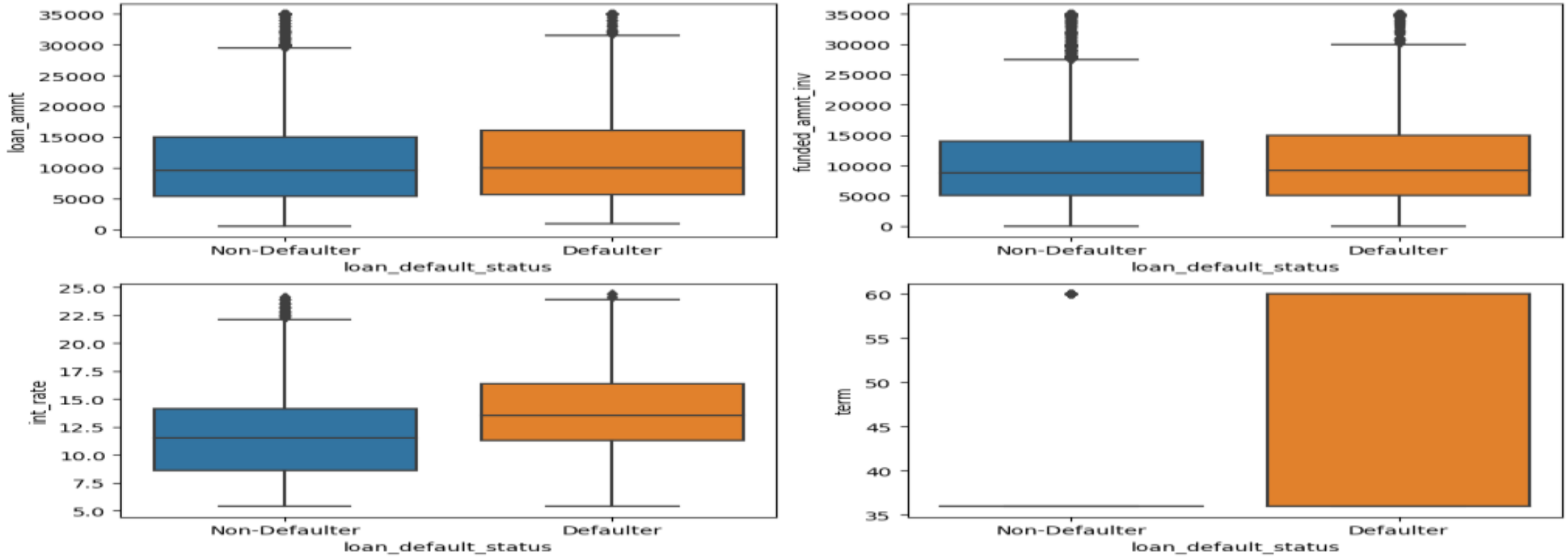
Inference: The state CA has most of the loans... The state CA has most of the 'Non-Defaulted' loans... The state CA has most of the 'Defaulted' loans... Conclusion: State CA tops in both Non-Defaulter and Defaulter catagories of loans.

Segmented analysis of Zip Code wise loan_status of state CA



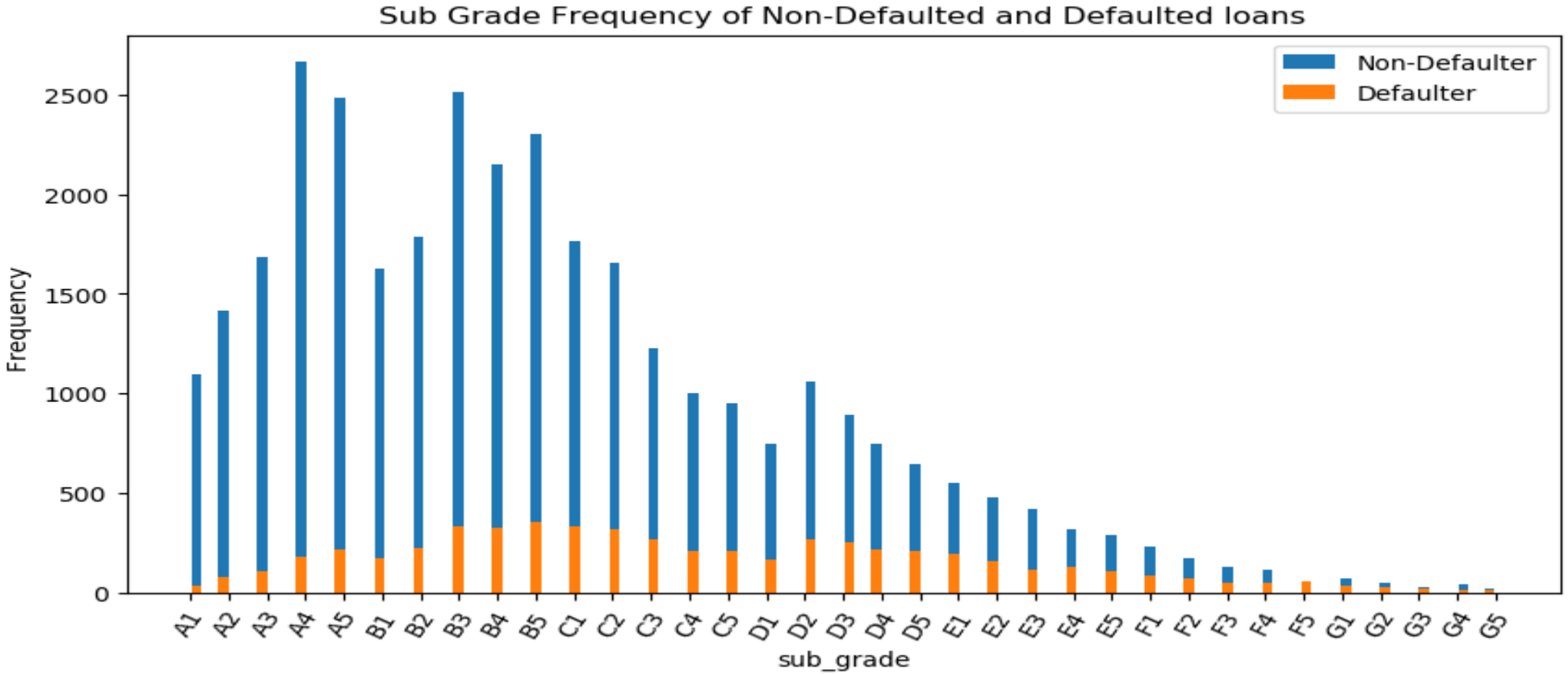
Inference: The zip code '945' in state CA has most of the loans.The zip code '945' in state CA has most of the Non-Defaulted loans... The zip code '945' in state CA has most of the Defaulted loans...
Conclusion: zip code '945' at State CA tops in Non-Defaulter and Defaulter categories of loans.

Segmented analysis based on loan status for loan amount, funded amount and interest rate



Inference: The loan_amnt and funded_amnt_inv has almost equal spread and median for defaulted and non-defaulted loans... The interest rate bracket for defaulted loans is clearly higher than non-defaulted loans... Almost All the loans Defaulted are given for 60 months interval... Conclusion: 60 months term loans and higher interest loans have maximum chances of defaulting

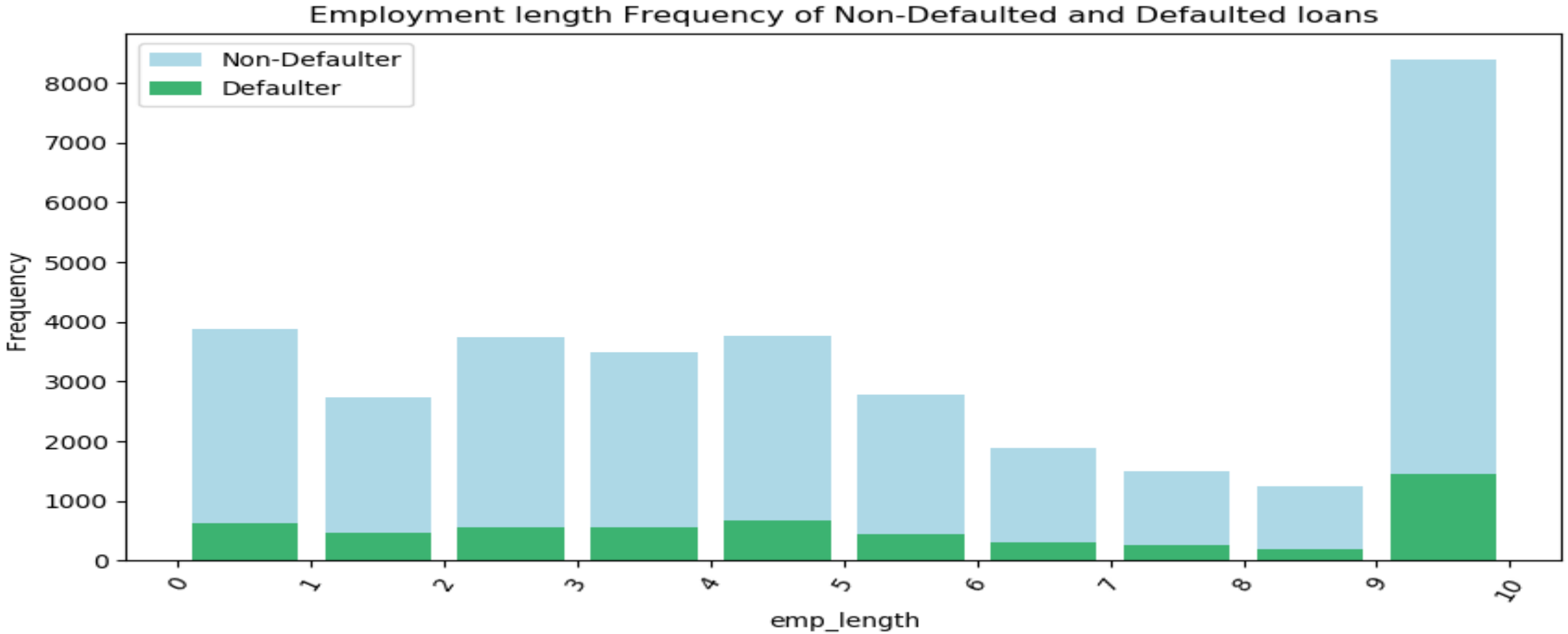
Segmented Analysis of sub grade



Inference:

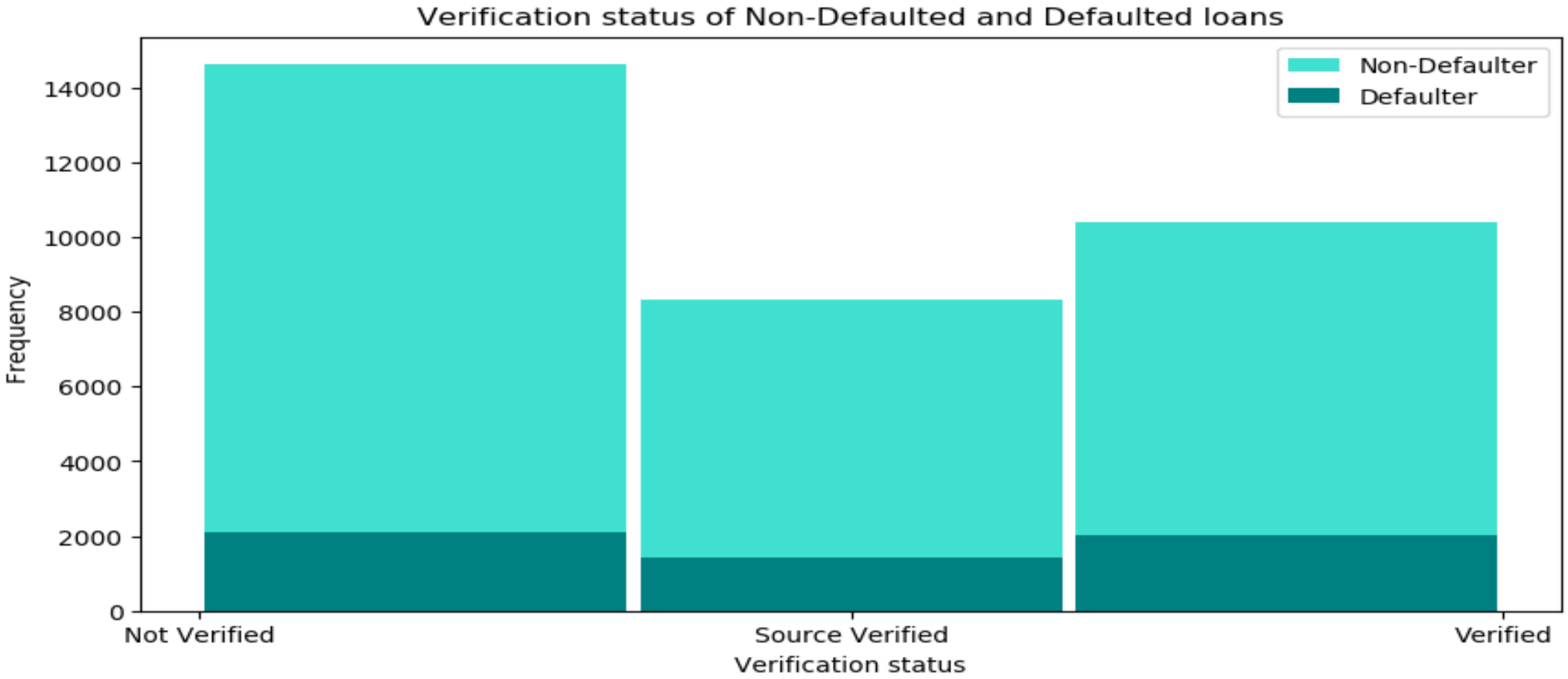
Lower the sub-grade higher is the chances of defaulting..

Segmented analysis of employment length



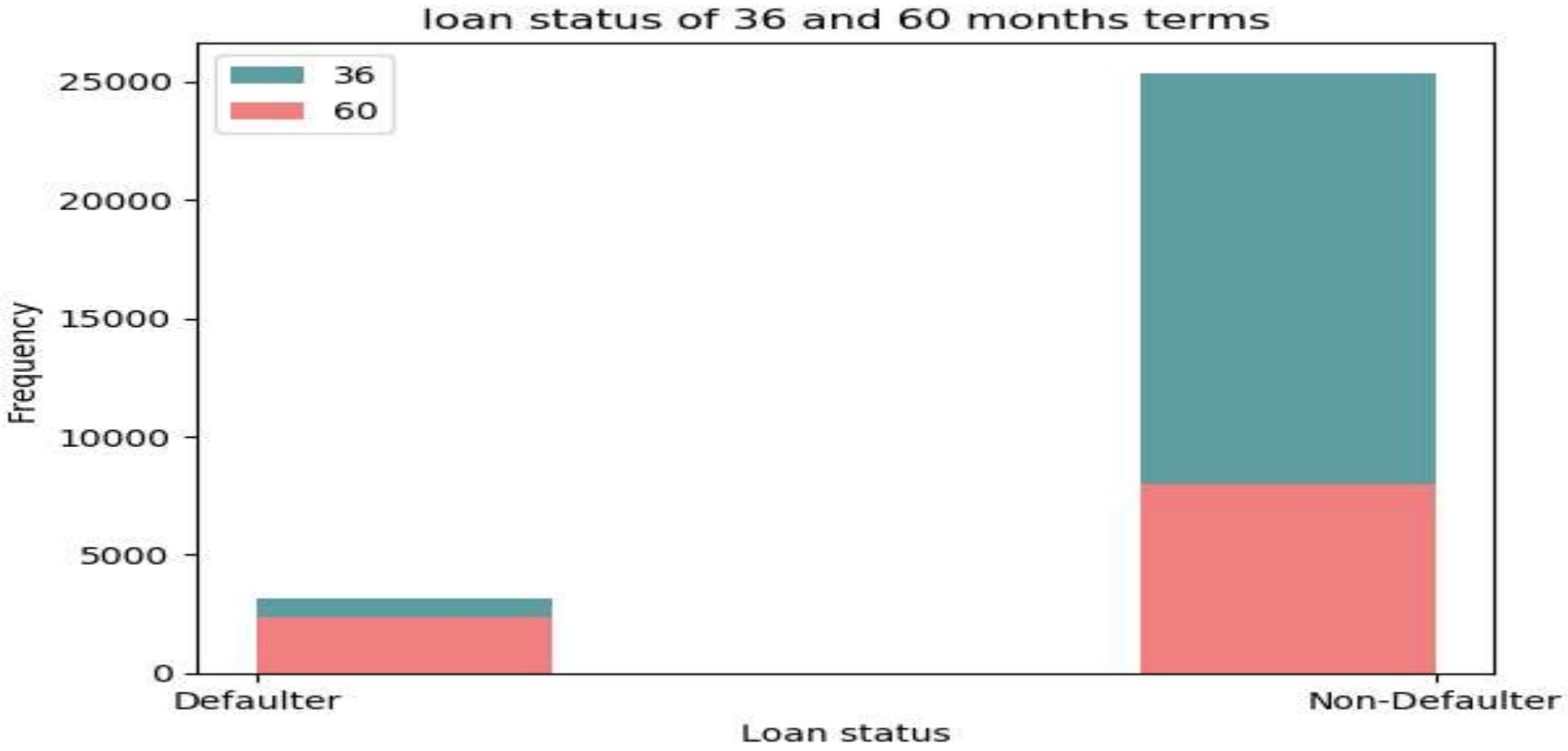
Inference: Lower experience employees tend to default more on loan payment than high experience employees

Segmented Analysis of Verification status



Inference: Maximum percentage(16%) of loans sanctioned in "Verified" category have been defaulted in loan while Not Verified have the least(12%)

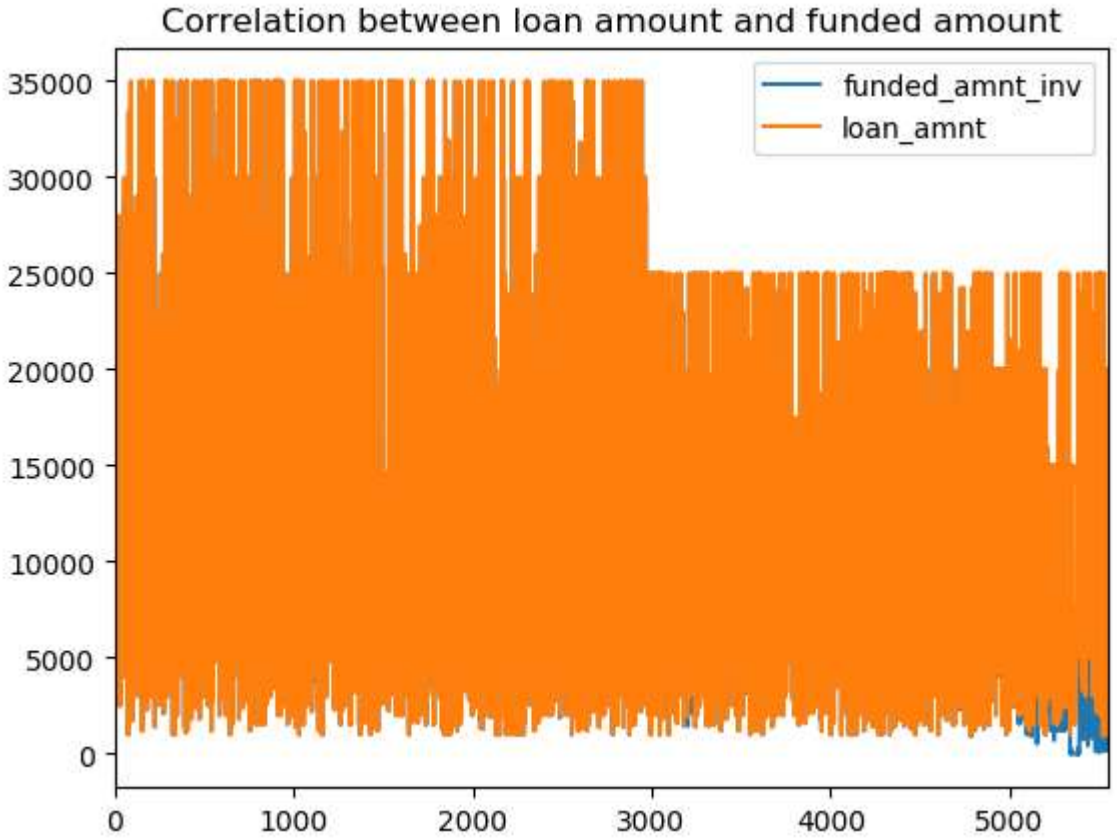
Segmented analysis of term

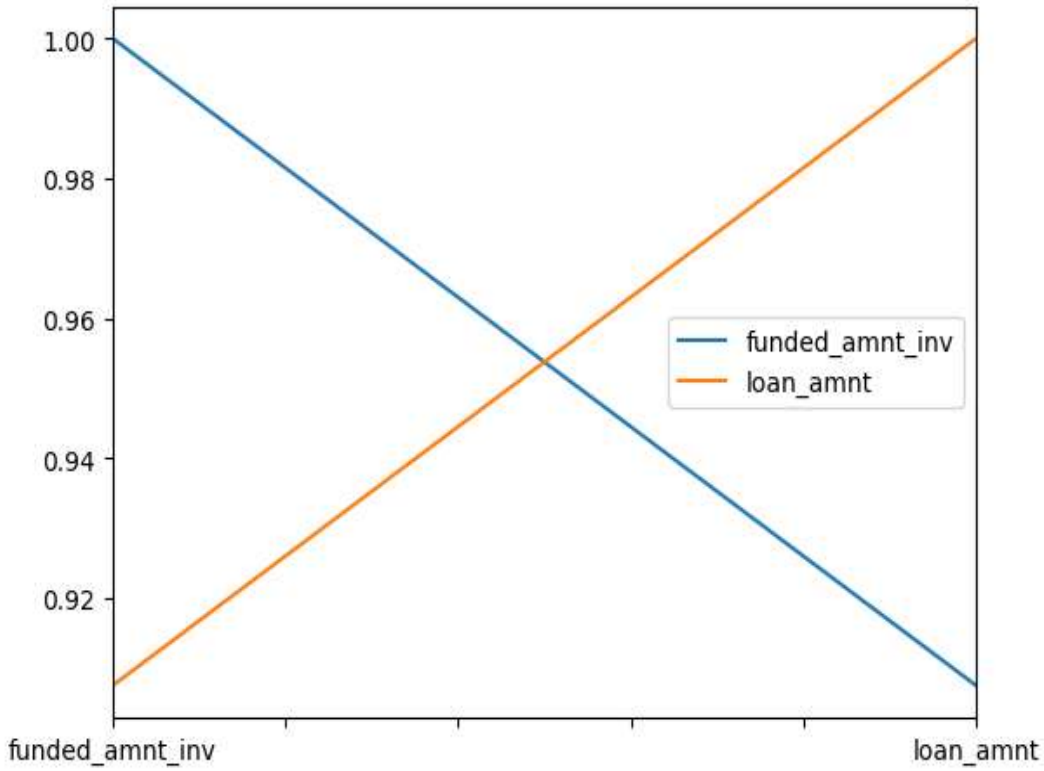


Inference:

The defaulted loans has mostly 60 months term.

Bivariate analysis on continuous variables

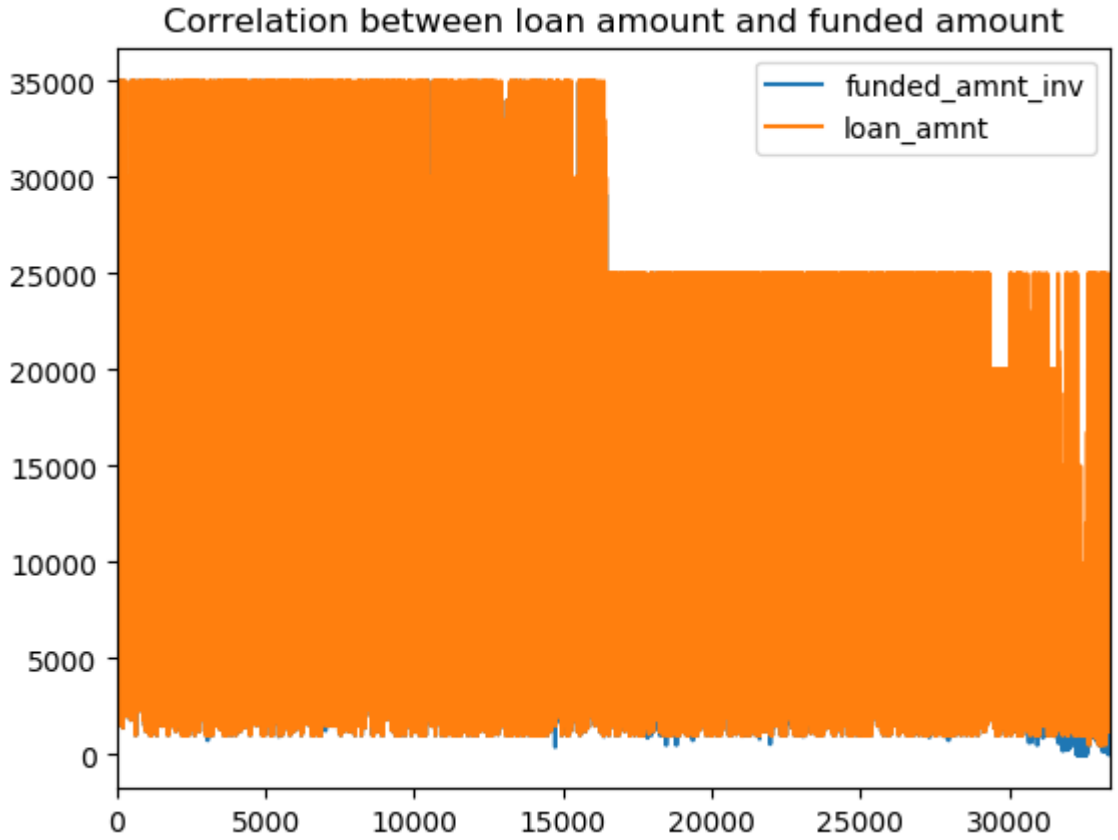


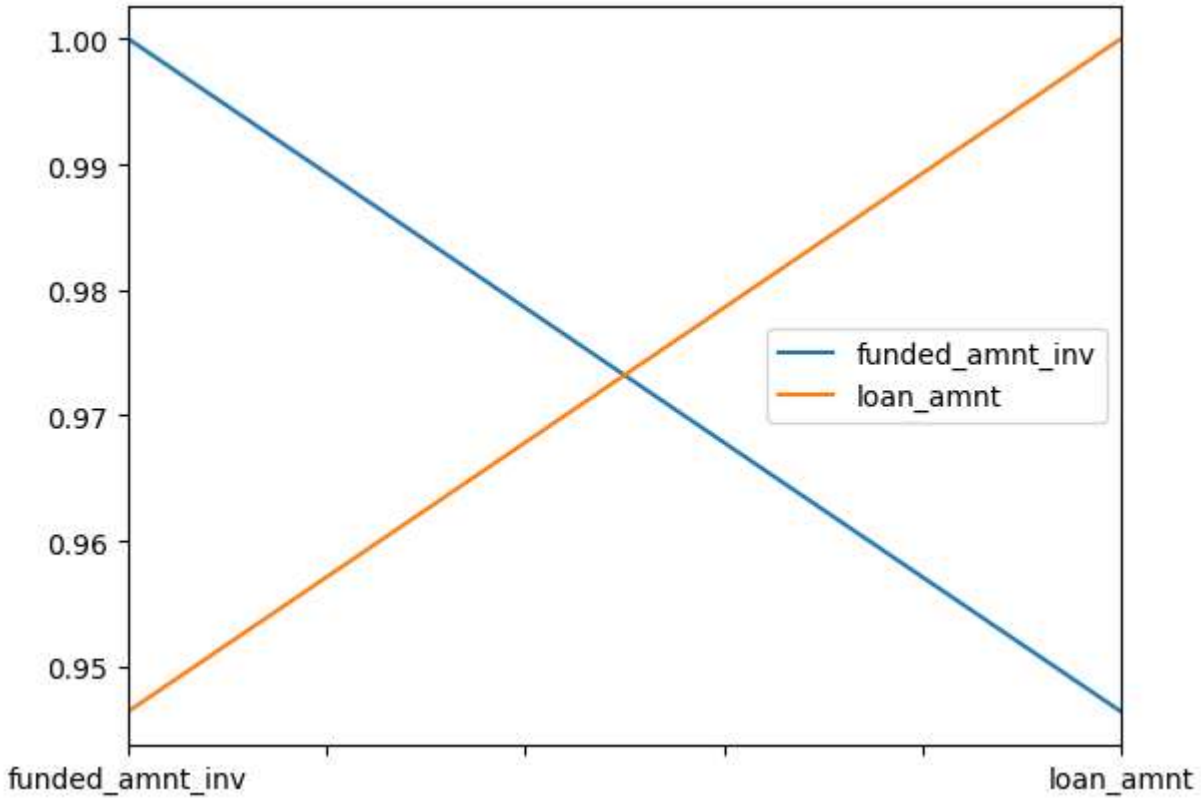


The correlation of loan amount and funded amount is 0.91

The loan amount and funded amount has higher correlation coefficient in the case of **Defaulted loans.**

Correlation of Loan amount and funded amount of Non-Defaulted loans

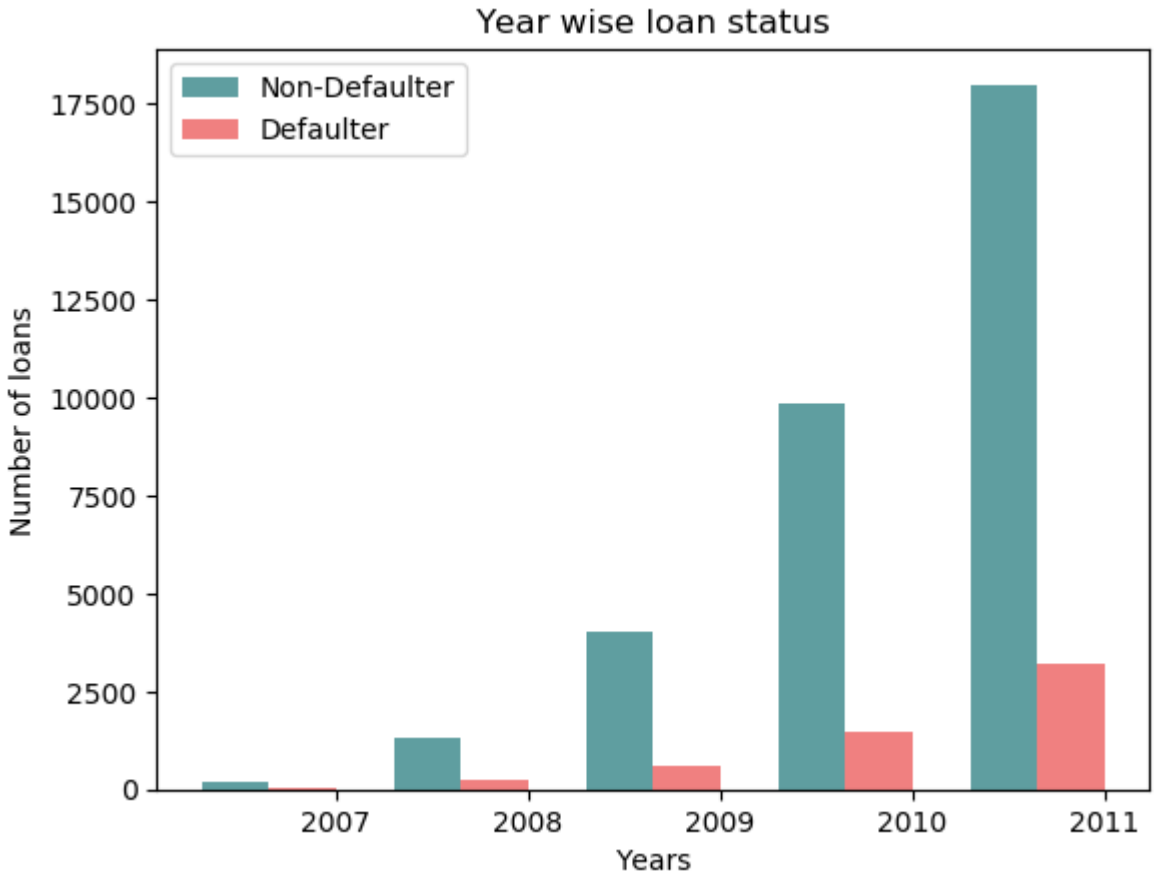


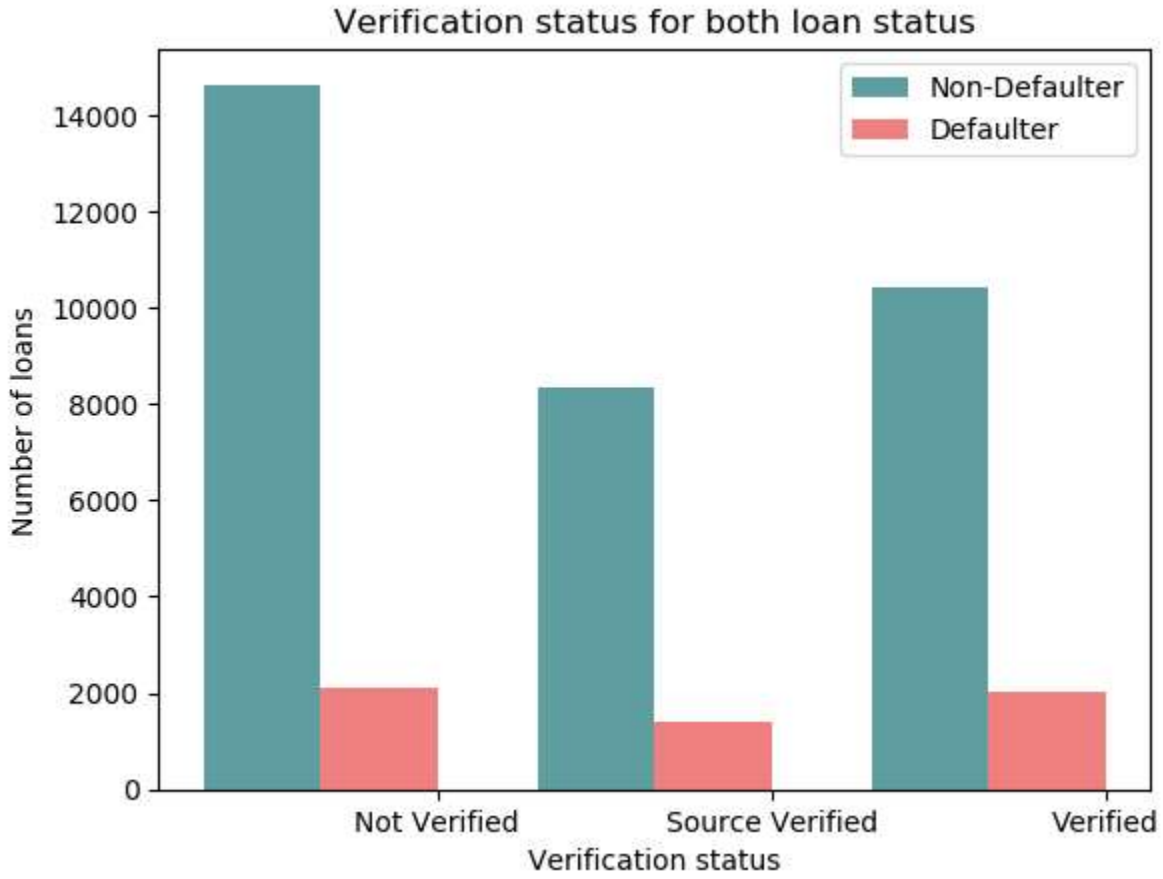


The correlation of loan amount and funded amount is 0.95

The loan amount and funded amount has higher correlation coefficient in the case of Non-Defaulted loans.

Bivariate analysis on categorical variables

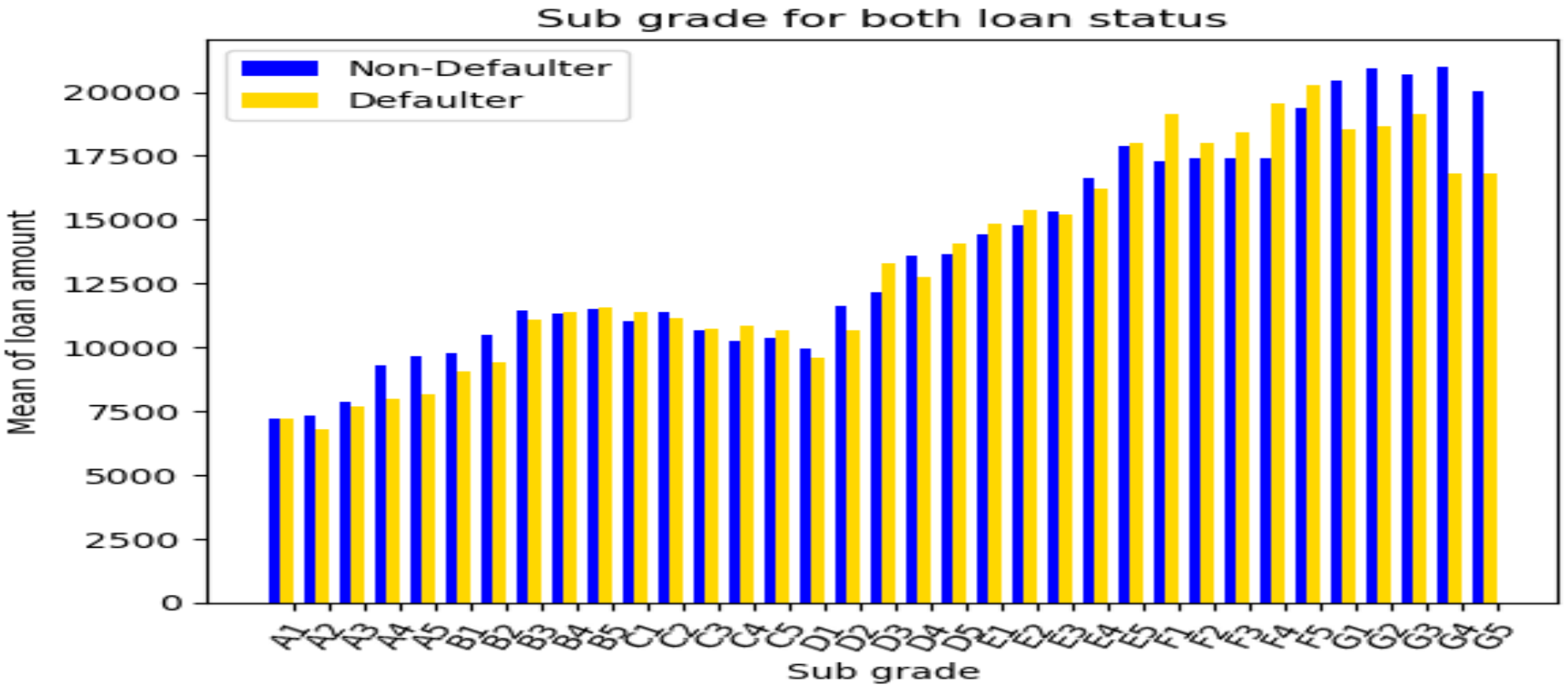




Inference:

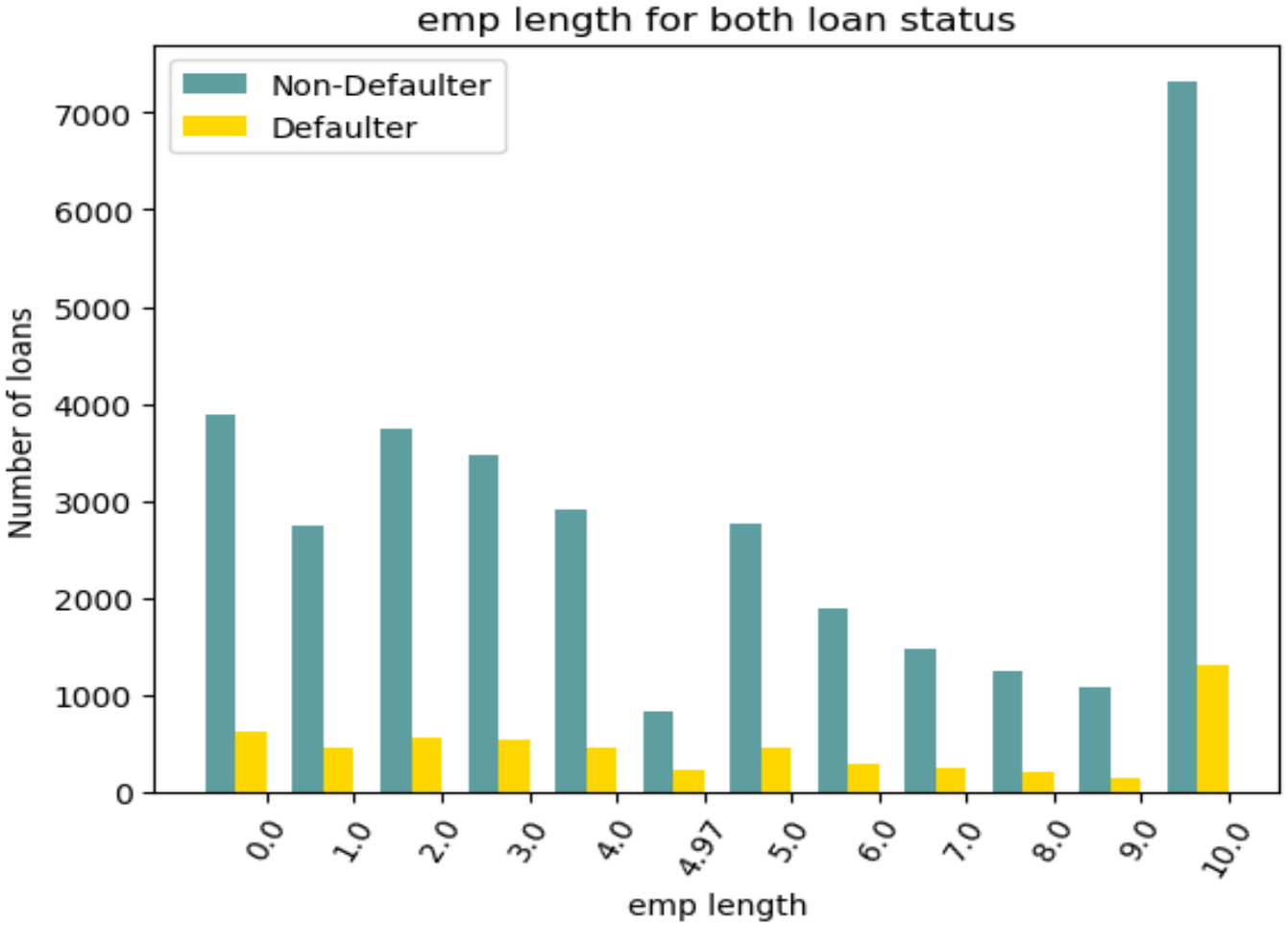
Year 2011 has most loans on both categories.

Analysis of Sub Grade and loan status



Inference: Lower the grade higher is the average loan amount and risk for defaulting except for few exceptions

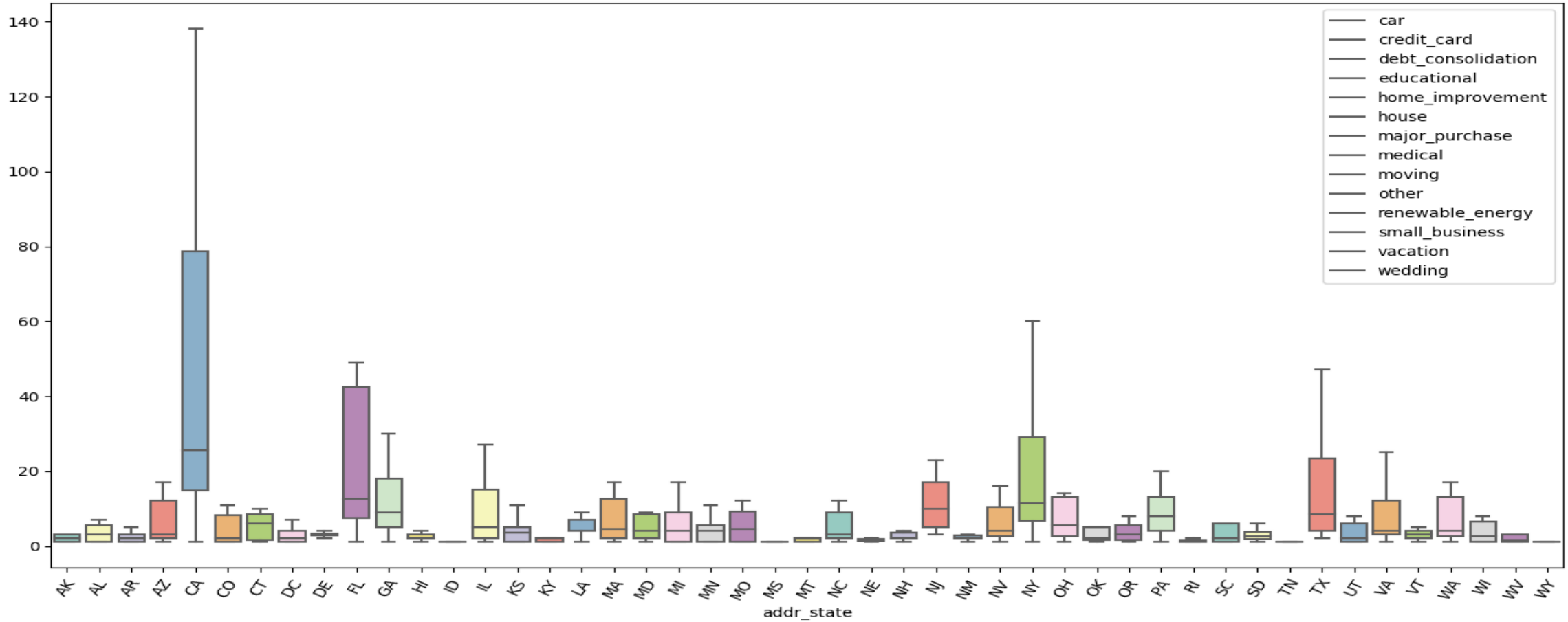
Analysis of Employment duration and loan status¶



Inference: Employees with longer duration are processing more loans in both categories.

There is no much deviation in defaulting rate based on employees' experience.

Analysis of state, purpose and loan status

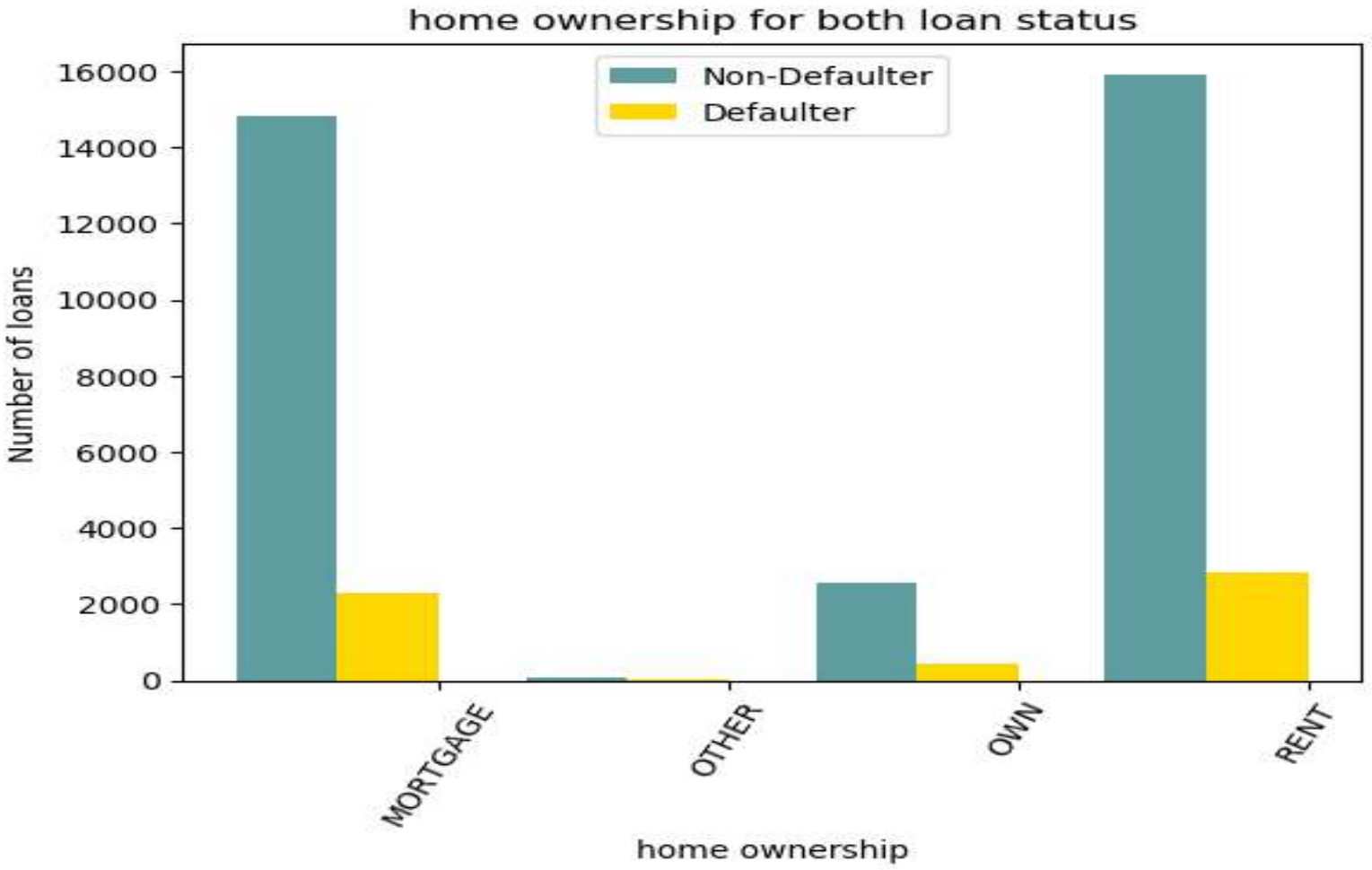


Inference:

The loan taken for debt consolidation has highest chances of defaulting

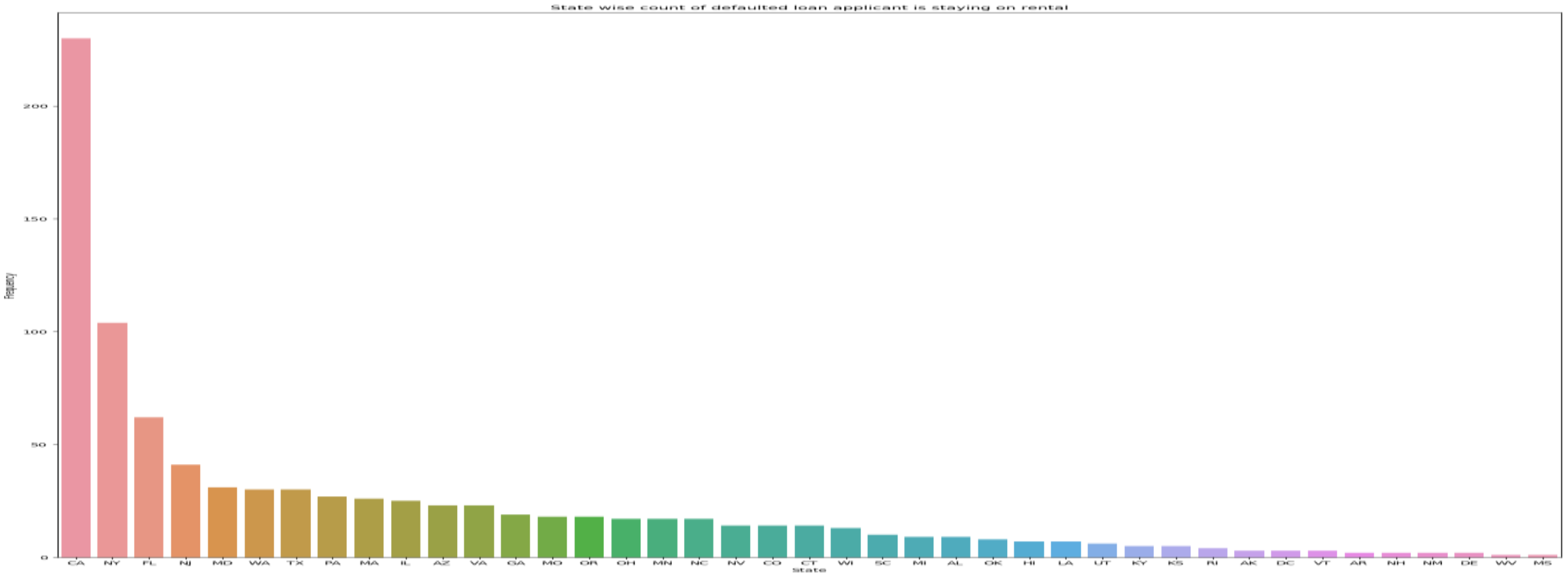
The state CA has highest percentage of defaulters

Analysis of home ownership and loan status



Inference: Applicant who are in rental homes are more likely to default followed by mortgage.

Analysis of address, verification status, home ownership and loan status



Inference: The State CA has maximum number of defaulted loan applicant staying on rental.

Analysis of verification status and loan status based on annual income

loan_default_status	Non-Defaulter
verification_status	
Not Verified	58669.808540
Source Verified	63014.910050
Verified	75343.250518

In each of the verification status group(Not verified, source verified, verified), Fully paid always has higher average annual_inc than charged off

loan_default_status	Defaulter		Non-Defaulter	
term	36	60	36	60
interest_rate_bracket				
10-15	6.594200	6.645357	6.132338	5.800984
15-20	7.258667	7.203643	7.100320	6.657143
20-25	9.570312	9.347769	7.218121	8.037000
5-10	5.415128	4.921428	4.951257	3.875707

For each term term loan the installment to income ration of for each of the interest slabs are as below. This influences whether a loan is likely to be defaulted or not:

Conclusion

- By analyzing the available data , we set up few decisive factors which helps in determining the loan defaulter applicant. So, the bank should consider these factors before approving the loan to avoid the credit losses.